Psychometrika

VOLUME XXIII—1958

JANUARY-DECEMBER

PUBLISHED QUARTERLY

By THE PSYCHOMETRIC SOCIETY
AT 1407 SHERWOOD AVENUE
RICHMOND 5, VIRGINIA

Psychometrika

CONTENTS

VOLUME TWENTY-THREE MARCH 1958 NUMBER 1

PSYCHOMETRIC MONOGRAPHS

The issues of this series are:

THURSTONE, L. L. Primary mental abilities.
Psychometric Monograph No. 1, $3.00. (Second impression, cloth binding.)

THURSTONE, L. L. AND THURSTONE, THELMA GWINN. Factorial studies of intelligence.
Psychometric Monograph No. 2, (out of print).

WOLFLE, DAEL. Factor analysis to 1940.
Psychometric Monograph No. 3, (out of print).

THURSTONE, L. L. A factorial study of perception.
Psychometric Monograph No. 4, (out of print).

FRENCH, JOHN W. The description of aptitude and achievement tests in terms of rotated factors.
Psychometric Monograph No. 5, $4.00.

DEGAN, JAMES W. Dimensions of functional psychosis.
Psychometric Monograph No. 6, $1.50.

LORD, FREDERIC. A theory of test scores.
Psychometric Monograph No. 7, $2.00.

ROFF, MERRILL. A factorial study of tests in the perceptual area.
Psychometric Monograph No. 8, $1.50.

Orders for Psychometric Monograph No. 1 should be sent to The University of Chicago Press, 5750 S. Ellis Avenue, Chicago 37, Illinois. Orders for No. 5 through No. 8 should be sent to The William Byrd Press, Box 2-w, Richmond 5, Virginia.

PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

GENERAL RESOLUTION OF CORRELATION MATRICES INTO COMPONENTS AND ITS UTILIZATION IN MULTIPLE AND PARTIAL REGRESSION*

JOHN A. CREAGER

AIR FORCE PERSONNEL AND TRAINING RESEARCH CENTER

The derivation of multiple and partial regression statistics from uniqueness-augmented factor loadings, presented in the literature for orthogonal factor solutions, is generalized to oblique solutions. A mathematical rationale for the general case, without restriction to uncorrelated factors, is presented. Use of the general formulation is illustrated with a two-factor, seven-variable example.

The considerable amount of computational time and labor required to compute multiple and partial correlation statistics when dealing with large test batteries is largely due to the necessity of computing the inverse of an nth order correlation matrix when classical procedures are used. Computation of multiple regression statistics from factor statistics permits considerable reduction in time and labor, especially when the number of variables is large and the number of factors is small [1, 3, 4]. Once the factorial reduction of the correlation matrix has been effected, any or all of the multiple and partial correlations or regression weights may be obtained. Furthermore, the factor solution may be studied to determine which predictors are most likely, when combined, to yield high prediction of a given variable.

The mathematical foundations and computational techniques for obtaining multiple and partial regression statistics have been presented for orthogonal factor solutions by Guttman [3], Guttman and Cohen [4], Dwyer [1], and Horst [5]. Some of the saving in computational effort is lost by the preliminary factor analysis, especially if the centroid method is used with computation of residuals after extracting each factor. Dwyer [2] has presented an example in which preliminary factoring was done using the square root or diagonal method. The multiple-group method, however, permits the extraction of several factors simultaneously and is therefore highly efficient. Since the multiple-group method will, in general, result in correlated factors, the solution must either be orthogonalized, which requires appreciable additional computation, or oblique factor statistics must be used directly to obtain the multiple and partial regression statistics.

It is the purpose of this paper to present the mathematical rationale,

*This report is based on work done under ARDC Project 7702, in support of the research and development program of the Air Force Personnel and Training Research Center, Lackland Air Force Base, Texas. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government.

and to demonstrate, by an illustrative example, the computational schemes for obtaining multiple and partial regression statistics from oblique factor solutions.

Fundamental Relations

Let R be an $n \times n$ correlation matrix of n variables with unit diagonals. Let R be factored, without restriction to uncorrelated factors, into r common factors and n unique factors, yielding

(*i*) a factor structure matrix, S, of order $n \times r$,

(*ii*) a factor intercorrelation matrix, ϕ, of order $r \times r$,

(*iii*) a factor pattern matrix, P, of order $n \times r$ obtained from $P = S\phi^{-1}$,

(*iv*) a diagonal matrix, U, of order n, giving the unique factor loadings.

Then

$$R = SP' + U^2. \tag{1}$$

Formula (1) states the fundamental factor theorem in general terms, where resolution of a correlation matrix is made into common factors, either correlated or uncorrelated, and unique factors which are uncorrelated either *inter se* or with the common factors.

In the subsequent development it is assumed that matrices R and U^2 are nonsingular. Let $V = U^{-1}$, and define $B = VS$ and $C' = P'V$, the uniqueness-augmented structure and pattern, respectively. Also let

$$Q = I + P'V^2S, \tag{2}$$

where Q is a Gramian matrix of order and rank r.

The Inverse of the Intercorrelation Matrix

The inverse of an intercorrelation matrix, R^{-1}, may be expressed in terms of oblique factor statistics. Starting with (1) and premultiplying both sides by $P'V^2$ gives

$$P'V^2R = P'V^2SP' + P' = (P'V^2S + I)\,P' = QP'. \tag{3}$$

Postmultiplying by R^{-1} gives

$$P'V^2 = QP'R^{-1}, \tag{4}$$

and therefore

$$Q^{-1}P'V^2 = P'R^{-1}. \tag{5}$$

Premultiplying both sides of (5) by S, the factor structure, and adding U^2R^{-1} gives

$$SQ^{-1}P'V^2 + U^2R^{-1} = I. \tag{6}$$

Subtracting $SQ^{-1}P'V^2$ from both sides and dividing by U^2 yields

$$R^{-1} = V^2(I - SQ^{-1}P'V^2) = V^2 - VBQ^{-1}C'V. \tag{7}$$

Use of (7) requires Q^{-1} which is of order r compared to R^{-1} which is order n.

Obtaining Regression Statistics

Standard regression weights to be applied to predictor variables in the multiple regression of a given criterion may be obtained in either of two ways. If partial correlation statistics are not required, the Q matrix may be developed by (2) using uniqueness-augmented factor statistics for the predictors only. Let this matrix be designated as Q_j , where j refers to the omitted criterion variable. If Q_j is used in (7), the inverse of the predictor intercorrelation matrix will be obtained. The desired regression weights may then be obtained by

$$\beta = R^{-1}r_c \, , \tag{8}$$

where r_c is a column vector of validity coefficients of order $n \times 1$, and β is a column vector of the desired weights. The multiple correlation coefficient for the set of predictors and the given criterion is given by

$$R_j^2 = \beta r_c' \, . \tag{9}$$

If regression weights are not required, the multiple correlation coefficient may be obtained directly from R^{-1} by

$$R_j^2 = \frac{R_{jj}^{-1} - 1}{R_{jj}^{-1}} . \tag{9a}$$

If partial correlations are desired, the inverse of the total correlation matrix, including the criterion validities, is required. In such a situation the regression weights and multiple correlation may be obtained from the Q matrix developed from the entire set of variables. The inverse, R^{-1}, is computed from the Q matrix as indicated by (7), the regression weights are then obtained by

$$\beta = -D^{-1}R^{-1}, \tag{10}$$

where D is a diagonal matrix derived from the diagonal elements of R^{-1}. The multiple correlation coefficients may then be computed as before by (9). Partial correlations holding constant $n - 2$ variables may be obtained by

$$R_{jk\cdot(n-2)} = -D^{-\frac{1}{2}}R^{-1}D^{-\frac{1}{2}}. \tag{11}$$

The Prediction of Factor Scores

The matrix of regression weights for predicting common factors from tests, W_c , is obtained from postmultiplying the inverse of the predictor intercorrelations by factor "validities" (the common factor structure),

$$W_c = R^{-1}S = V^2(I - SQ^{-1}P'V^2)S = V^2S - VBQ^{-1}C'VS. \tag{12}$$

Similarly, the matrix of regression weights for predicting unique factor scores, W_u, is

$$W_u = R^{-1}U = [V^2 - VBQ^{-1}C'V]\,U = V - VBQ^{-1}C'. \tag{13}$$

The corresponding squared multiple correlation coefficients may then be obtained as the product sum of regression weights and validities.

In a situation in which only the multiple correlation coefficient for predicting a common factor from test scores is desired, and the regression weights are not needed for a prediction equation, it may be obtained very readily without computation of R^{-1} or the regression weights. The multiple correlation coefficient for a common factor from tests and the remaining common factors is equal to that from tests alone, since all of the common variance is in the test battery and adding the common variance to the battery will not change its predictive power. Guttman [3] and Dwyer [1] have shown that the multiple correlation coefficient for predicting a common factor from remaining factors and tests, for the orthogonal case, is

$$R_f = \sqrt{1 - \frac{1}{1 + \sum_{j=1}^{n} B_{jf}^2}} = \sqrt{\frac{\sum B_{jf}^2}{1 + \sum B_{jf}^2}}. \tag{14}$$

A similar development for oblique factors yields

$$R_f = \sqrt{\frac{\sum_{j=1}^{n} B_{jf}C'_{jf}}{1 + \sum_{j=1}^{n} B_{jf}C'_{jf}}}. \tag{15}$$

Computational Techniques

To illustrate computational techniques for the application of the principles developed above, the seven-variable, two-factor example used by Dwyer [1] is convenient, although the saving in computational effort becomes more convincing as the number of tests increases more rapidly than the number of factors. The correlation matrix is given in Table 1 with exact communalities in the diagonal cells. This matrix was factored by the multiple-group method, the summations being made over variables 1, 2, and 7 for factor I, and over variables 3 and 4 for factor II. The resulting factorial statistics are shown in Table 2. In usual applications where exact communalities are not known, it is necessary to use estimates [7].

In a practical situation it is necessary to judge the rank of R and to test this judgment by examination of the residuals. If r is underestimated, appreciable residuals will remain; if r is overestimated, some of the saving in computational labor will be lost. It is essential that residuals be negligible before proceeding with computation of regression statistics. Otherwise the

TABLE 1

The Reduced Correlation Matrix*

Test	R 1	2	3	4	5	6	7
1	450	580	-280	010	360	380	610
2	580	760	-280	100	520	440	780
3	-280	-280	700	560	140	-560	-420
4	010	100	560	610	400	-340	-030
5	360	520	140	400	540	080	460
6	380	440	-560	-340	080	520	540
7	610	730	-420	-030	460	540	830

*Decimal points have been omitted.

latter will be approximated to a degree dependent upon the magnitude of residuals. The multiple correlation obtained under these conditions will generally be high by an amount approximately equal to the average of the absolute residual error [1].

Once the factorial reduction of R has been accomplished and the rth residuals checked for an indication of the completeness of extraction, the diagonal matrices V^2 and V are computed by taking the reciprocals of U^2 and U, respectively. Each row of the factor structure and pattern is then multiplied by v_{ii} to obtain the uniqueness-augmented structure, B, and the uniqueness-augmented pattern, C. These are shown in Table 3.

The next step is forming the matrix Q. This is done by summing uniqueness-augmented, structure-pattern cross products as follows:

$$(16) \qquad Q = \begin{bmatrix} 1 + \sum B_{i\mathrm{I}}C_{i\mathrm{I}} & \cdots & \sum B_{i\mathrm{II}}C_{i\mathrm{I}} & \cdots & \sum B_{ir}C_{i\mathrm{I}} \\ \sum B_{i\mathrm{I}}C_{i\mathrm{II}} & & & & \vdots \\ \vdots & & & & \vdots \\ \sum B_{i\mathrm{I}}C_{ir} & & & & 1 + \sum B_{ir}C_{ir} \end{bmatrix}.$$

The summations are performed across tests, including the criterion variable, and across whatever predictor variables one may wish to include in the prediction. The Q matrix for all seven variables is shown for the illustrative example in Table 3. It is important to remember to add unity to the cross-product summations for the diagonal values of Q.

Table 4 shows the methods outlined for predicting variable 1. Matrices B and C' were obtained from Table 3 and matrix Q^{-1} by inversion of the Q matrix (involving all seven variables) in Table 3. The subsequent operations are also illustrated using variable 1 as the criterion variable. It is seen that, in the usual practical situation, only single rows of the subsequent matrices need to be computed. Hence, only the first row of each of the subsequent

TABLE 2

The Factor Statistics*

Test	S		P		Communality	Uniqueness
	I	II	I	II	h^2	u^2
1	6706	-1732	6669	-0158	4500	5500
2	8669	-1155	8892	0944	7600	2400
3	-4008	8083	-2224	7558	7000	3000
4	0327	7506	2223	8031	6100	3900
5	5480	3464	6670	5038	5400	4600
6	5561	-5774	4446	-4724	5200	4800
7	9078	-2887	8892	-0788	8300	1700

	ϕ			ϕ^{-1}	
	I	II		I	II
I	1.0000	-.2361	I	1.0590	.2500
II	-.2361	1.0000	II	.2500	1.0590

*Decimal points have been omitted except in ϕ and ϕ^{-1}.

TABLE 3

Uniqueness-augmentation of the Factor Statistics

Test	v^2_{JJ}	v_{JJ}	B		C	
			B_{JI}	B_{JII}	C_{JI}	C_{JII}
1	1.8182	1.3484	0.9043	-0.2335	0.8993	-0.0213
2	4.1665	2.0412	1.7695	-0.2358	1.8151	-0.1927
3	3.3335	1.8258	-0.7318	1.4758	-0.4061	1.3800
4	2.5642	1.6013	0.0524	1.2019	0.3560	1.2860
5	2.1742	1.4745	0.8080	0.5108	0.9835	0.7428
6	2.0834	1.4434	0.8027	-0.8334	0.6417	-0.6819
7	5.8826	2.4254	2.2018	-0.7002	2.1567	-0.1911

	Q			Q_1	
	I	II		I	II
I	11.3993	-2.3520	I	10.5861	-2.1420
II	-0.9887	5.6233	II	-0.9694	5.6183

TABLE 4

Matric Computations for Regression Statistics

	B			Q^{-1}		BQ^{-1}	
	I	II		I	II	I	II
1	0.9043	-0.2335	I	0.09103	0.03807	0.0786	-0.0086
2	1.7695	-0.2358	II	0.01600	0.18453	0.1573	0.0239
3	-0.7318	1.4758				-0.0430	0.2447
4	0.0524	1.2019				0.0240	0.2238
5	0.8080	0.5108				0.0817	0.1250
6	0.8027	-0.8334				0.0597	-0.1232
7	2.2018	-0.7002				0.1892	-0.0454

	C'						
	1	2	3	4	5	6	7
I	0.8993	1.8151	-0.4061	0.3560	0.9835	0.6417	2.1567
II	-0.0213	0.1927	1.3800	1.2860	0.7428	-0.6819	-0.1911

$BQ^{-1}C'$(row 1)

0.0709	0.1410	-0.0438	0.0169	0.0709	0.0563	0.1712

VV' (row 1)

1.8182	2.7524	2.4619	2.1592	1.9882	1.9463	3.2704

R^{-1} (row 1)

1.6893	-0.3881	0.1078	-0.0365	-0.1410	-0.1096	-0.5599

$d_1 = 1.6893$ $\sqrt{d_1} = 1.2997$ $1/d_1 = 0.59196$ $1/\sqrt{d_1} = 0.76941$

β_{1k} (row 1)

——	0.2297	-0.0638	0.0216	0.0835	0.0649	0.3314

matrices is shown in Table 4, row 1 of VV' was obtained by multiplying each v_{jj} by 1.3484 (v_{11} from Table 3). The first row of R^{-1} is then obtained multiplying each element of the row of $BQ^{-1}C'$ by the corresponding element of the same row of VV' and reversing the sign. The jth element (i.e., the diagonal element) of the row (in this case, the first cell entry) is then adjusted by adding v^2_{jj} from Table 3. Thus the 1.6893 in the first cell of R^{-1} in Table 4 was obtained by multiplying $0.0709 \times (-1.8182) = -0.1289$ and adding 1.8182.

The regression coefficients for predicting variable 1 are then obtained by multiplying each element in the row of R^{-1} by $-1/d_i$, where d_i is the diagonal element of R^{-1}.

The inverse of R may be checked by recalling that $RR^{-1} = I$. In the present example the first row of R multiplied by the first row of R^{-1} gives 0.9997, and the second row of R multiplied by the first row of R^{-1} gives $-$ 0.0004. It is, of course, the complete correlation matrix and its inverse that is involved here, rather than the reduced matrix shown in Table 1.

The square of the multiple correlation of variable 1 in the other six variables is obtained by multiplying the first row of the β matrix by the first row of R, omitting R_{11} . This gives $R^2_{1\cdot 2 \ldots\ldots 7} = 0.408182$ and $R_{1.234567} =$ .6389. Use of formula (9a) gives $R^2_{1\cdot 2 \ldots\ldots 7} = 0.408039$ and $R_{1.234567} = .6388$, the value obtained by Dwyer. β_{12} is $0.3881 \times 0.59196 = .2297$.

To obtain the partial correlation between variables 1 and 2 holding constant the remaining five variables, the diagonal element of the second row of R^{-1} is required. The corresponding element of $BQ^{-1}C'$ is (0.1573) (1.8151) + (0.0239) $\times$ (0.1927) = 0.2901; $v^2_{jj} = 4.1665$. The negative product of these is $-$ 1.2087, and $d_2 = 2.9578$. The partial correlation coefficient is then obtained from the (1, 2) cell of R^{-1}.

$$-1/\sqrt{d_1}\cdot 1/\sqrt{d_2} = -0.3881 \times -0.7694 \times 0.5815 = 0.1736.$$

By similar operations, applying (12) and (13), regression statistics for the prediction of factor scores may be obtained.

Discussion

The methods of regression analysis from uniqueness-augmented factor statistics given by Dwyer [1] are formulated in terms of determinants. Generalization of Dwyer's method is possible for the oblique factor statistics. Both Dwyer's method and the one presented here in matrix terms are readily adapted to machine methods of statistics. By having either method in terms of oblique factor statistics, multiple-group extraction methods may be used to minimize residual computations without requiring orthogonalization of the factor matrices.

These techniques are useful when it is desired to obtain: (*i*) the regression of each variable on the $n - 1$ remaining variables; (*ii*) the partial regression of each pair of variables, holding constant the remaining $n - 2$ variables; (*iii*) the regression weights for the prediction of test scores; (*iv*) the regression weights for the prediction of factor scores. They can also be used to set up standard procedures for routine treatment of batteries by machine methods of statistics.

REFERENCES

[1] Dwyer, P. S. The evaluation of multiple and partial correlation coefficients from the factorial matrix. *Psychometrika*, 1940, **5**, 211-232.

[2] Dwyer, P. S. The relative efficacy and economy of various test selection methods. PRS Report 957, AGO. 12 June 1952.

[3] Guttman, L. Multiple rectilinear prediction and the resolution into components. *Psychometrika*, 1940, **5**, 75-99.

[4] Guttman, L. and Cohen, J. Multiple rectilinear prediction and the resolution into components: II. *Psychometrika*, 1943, **8**, 169-183.

[5] Horst. P. (Ed.) The prediction of personal adjustment. *SSRC Bull.*, 1941, **48**, pp. 437ff.

[6] Thorndike, R. L. *Personnel selection.* New York: Wiley, 1949.

[7] Thurstone, L. L. *Multiple-factor analysis.* Chicago: Univ. Chicago Press, 1947.

Manuscript received 7/16/56
Revised manuscript received 5/23/57

PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

ERROR OF MEASUREMENT AND THE SENSITIVITY OF A TEST OF SIGNIFICANCE

J. P. SUTCLIFFE*

UNIVERSITY OF SYDNEY

Implications of random error of measurement for the sensitivity of the *F* test of differences between means are elaborated. By considering the mathematical models appropriate to design situations involving true and fallible measures, it is shown how measurement error decreases the sensitivity of a test of significance. A method of reducing such loss of sensitivity is described and recommended for general practice.

In the statistical theory of sampling, explicit attention is given to sampling error, which refers to fluctuations in the composition of samples drawn at random from a defined universe. A second form of error, largely ignored in this context, is measurement error. This applies to the individual sampling units and is thus related to the definition of the universe rather than sampling outcomes. Applications of sampling theory have proceeded on the implicit assumption that the sampling units which make up the defined universe are error free, that (in psychometric terms) the universe consists of *true* scores. This assumption is not justified in practice, where measurement is seldom free from error. Parameters, such as the mean and the variance, of a universe of *fallible* scores will differ from those of a universe of true scores; tests of significance of a given effect will not necessarily be the same in the two cases. This paper elaborates the implications of measurement error for the simple case of the *F* test of difference between means. By setting up the mathematical models appropriate to the relevant design situations, it is shown how measurement error (relative to the parallel true score case) decreases the sensitivity of the test of significance. Sensitivity refers to the likelihood of detecting a nonzero population effect at a given level of significance. Through its inverse, proneness to Type II error, it is usually expressed quantitatively as power. A method of reducing such loss of sensitivity is described.

Definition of Universes of Scores

The scale or range of application of a measuring instrument comprises a number of units of measurement. Let w represent any one unit or subrange of the scale and v any one occasion of measurement. Errors of measurement

*I wish to express my thanks in acknowledgement that the present form of this paper has benefited from editorial comment, and from the advice of Dr. H. Mulhall of the Department of Mathematics, University of Sydney.

constant for all units of the scale on all occasions of testing will be designated f; errors constant for all occasions of measurement with a particular unit, but variable from unit to unit will be designated g_w ; errors variable from occasion to occasion and from unit to unit will be designated h_{vw} . For example, a carpenter's tape may be incorrectly calibrated uniformly over the whole scale; then unevenly stretched over the first few feet which are most commonly used; and finally subject to random error on any given application. For this case the total error of measurement $E = f + g_w + h_{vw}$. Analogous errors of measurement occur with psychological tests [3], but these will not be discussed here; while knowledge of the source of error can facilitate its control, it is rather the mode of operation of error which is relevant to the statistical argument.

Most generally, an obtained fallible measure or score, X_v , can be expressed as the sum of the true score, T_v , and its error of measurement, E_v [3]. This holds whether measurement error is unitary, or complex in the sense illustrated above. The additive relationship also holds whatever *other* relationship may be shown to obtain between true score and error for a universe of obtained scores. For instance, while E'_v may enter as a multiplier in the relationship between obtained and true score, $X_v = E'_v T_v$, X_v may also be written $X_v = T_v + E_v$, where $E_v = (E'_v - 1)T_v$. Other assumptions about the nature of error and its relationship to true score are tenable, but the additive assumption is adopted here because it simplifies the subsequent analysis.

The mean and variance of an infinite universe of fallible scores $X_v = T_v + E_v$ may be obtained as follows:

$$\text{Mean} = \lim_{N \to \infty} [\sum^{N} X_v/N] = \lim_{N \to \infty} [\sum^{N} (T_v + E_v)/N] = \bar{T} + \bar{E}.$$

$$\text{Variance} = \lim_{N \to \infty} [\sum^{N} x_v^2/N] = \lim_{N \to \infty} [\sum^{N} (t_v + e_v)^2/N]$$

$$= \sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e .$$

These outcomes are summarized in Table 1. Depending upon the mode of operation of error, cases may arise where any or all of $\bar{E}$, σ_e^2 , and ρ_{te} are zero,

TABLE 1

Parameters of Universes of True, Error and Obtained Scores

Universe	Mean	Variance
True scores T_v	$\bar{T}$	σ_t^2
Error scores E_v	$\bar{E}$	σ_e^2
Obtained scores X_v	$\bar{T} + \bar{E}$	$\sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$

in which cases one or more of the parameters will be common to the universes of true and obtained scores.

When error is absent, the mean $= \bar{T}$ and variance $= \sigma_t^2$ (Case 1). When error is constant $\bar{E} = f > 0$, $\sigma_e^2 = 0$, $\rho_{te} = 0$; hence the mean of fallible scores $= \bar{T} + f$, and variance $= \sigma_t^2$ (Case 2). Where error is variable its distribution may be either random or nonrandom. (In either case, the variances of error about different true score values may be homogeneous or heterogeneous. Heterogeneity of variance permits nonzero correlation between true scores and the *variance* of errors about them, but, as in random sampling, this correlation is independent of ρ_{te}. Heterogeneity of error variance should, of course, be taken into account in any analysis of variance [2].) If errors occur at random about T_v, then $\bar{E} = 0$, $\sigma_e^2 > 0$, and $\rho_{te} = 0$; hence mean $= \bar{T}$, and variance $= \sigma_t^2 + \sigma_e^2$ (Case 3). If errors are randomly distributed about $T_v + f$, $\bar{E} = f + 0$, $\sigma_e^2 > 0$, $\rho_{te} = 0$; hence mean $= \bar{T} + f$, and variance $= \sigma_t^2 + \sigma_e^2$ (Case 4). Where errors are distributed randomly about $T_v + g_w$, then $\bar{E} = \bar{g} + 0$, $\sigma_e^2 > 0$, $\rho_{te} > 0$, and hence mean $= \bar{T} + \bar{g}$, and variance $= \sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$ (Case 5). With nonrandom distribution of errors, generally one would find $\bar{E} > 0$, $\sigma_e^2 > 0$, and $\rho_{te} > 0$. Whether errors are distributed about T_v, $T_v + f$, or $T_v + g_w$, mean $= \bar{T} +$ error, and variance $= \sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$. All cases of nonrandom distribution of error are here referred to as Case 6.

The six cases are summarized in Table 2 to enable comparison of the parameters of fallible score universes with those of the true score universe. In no case are both parameters the same as those in Case 1; however, Case 2 has the same variance, and Case 3 the same mean. Cases 1 and 2 are unlikely to occur in practice. Most experiments aim to achieve the conditions of Case 3, but the intrusion of constant errors, scale biases, and other nonrandom errors makes Cases 4, 5, and 6 quite common. The following discussion will center on Cases 1 and 3, with incidental comment on the others.

TABLE 2

Parameters of Universes of True and Fallible Scores

Case	Mean	Variance
1	$\bar{T}$	σ_t^2
2	$\bar{T} + f$	σ_t^2
3	$\bar{T}$	$\sigma_t^2 + \sigma_e^2$
4	$\bar{T} + f$	$\sigma_t^2 + \sigma_e^2$
5	$\bar{T} + \bar{g}$	$\sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$
6	$\bar{T}$ + error	$\sigma_t^2 + \sigma_e^2 + 2\rho_{te}\sigma_t\sigma_e$

Comparison of the Design Models

With the universes of true and fallible scores defined, it becomes possible to compare the sensitivity of tests of significance applied in given cases. For comparative purposes the analysis of variance for Case 1 will be described. Then two analyses for Case 3 will be considered—the first reflecting common practice, the second involving random replication of measurement to increase reliability and hence sensitivity.

Notation and plan for Case 1

Consider the comparison of means of independent random samples of true scores obtained at different levels of a single-treatment classification. Let $i = 1, 2, \cdots, a$ represent any one of the treatment levels within the treatment classification A. Let $j = 1, 2, \cdots, b$ represent any one subject in a sample of subjects B. Then X_{ij} is the true score of the subject j in the treatment level or group i. As subjects are randomly sampled, j represents number only, not rank within a group. Let a dot in place of a subscript represent summation across the class indicated by the subscript replaced, e.g.,

$$\sum_{j=1}^{b} X_{ij} = X_{i.}\,, \qquad \sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij} = X_{..}\,.$$

The sample values of X_{ij} and the sums are represented in Table 3.

TABLE 3

Plan of Obtained Scores of Subjects Within Random Samples Allocated to Independent Treatment Groups

A				B Subjects			
Treatments	1	2	.	j	.	b	Sum
1	X_{11}	X_{12}	.	X_{1j}	.	X_{1b}	$X_{1.}$
2	X_{21}	X_{22}	.	X_{2j}	.	X_{2b}	$X_{2.}$
.	.	.	.	.	.	.	.
i	X_{i1}	X_{i2}	.	X_{ij}	.	X_{ib}	$X_{i.}$
.	.	.	. .	.	.	.	.
a	X_{a1}	X_{a2}	.	X_{aj}	.	X_{ab}	$X_{a.}$
							$X_{..}$

Analysis of variance for Case 1

The total variance of the ab sample values of X_{ij} can be expressed in terms of two sources of variation: between treatments, A, and between subjects within treatment levels, B_A. A given deviation score may be written as

$$x_{ij} = (X_{ij} - \bar{X}_{..}) = (\bar{X}_{i.} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.}).$$

The total sum of squares is

$$\mathrm{SS}_T = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{..})^2 = b\sum_{i=1}^{a}(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{i.})^2.$$

The degrees of freedom pertaining to these components are Total = $(ab - 1)$, $A = (a - 1)$, $B_A = a(b - 1)$. From the SS and df, the mean squares, S^2, may be obtained as unbiased estimates (on the null hypothesis) of a common population variance.

Expectation of mean squares for Case 1

To determine what is estimated by a given S^2, one takes the expectation according to the model involved. As Case 1 involves a universe of true scores, Model 1 can be written as

$$X_{ij} = A_i + B_{ij} .$$

A_i is the class of treatment parameters of which the sampled treatment means are estimators. The distribution of A_i will vary according as treatments are fixed constants or randomly sampled. For convenience the case of random A_i with variance σ_A^2 will be considered. B_{ij} is the class of true score deviations from A_i , which are normally distributed with zero mean and variance σ_t^2 . To find the expected values of SS and then S^2, one substitutes model values in the analysis of sample variance and thereby determines the limiting value of a given component.

(*i*) *Expectation of* S_A^2

$$(\bar{X}_{i.} - \bar{X}_{..}) = (A_i - \bar{A}_{.}) + (\bar{B}_{i.} - \bar{B}_{..});$$

$$E\left\{b\sum_{i=1}^{a}(\bar{X}_{i.} - \bar{X}_{..})^2\right\} = E\left\{b\sum_{i=1}^{a}(A_i - \bar{A}_{.})^2 + b\sum_{i=1}^{a}(\bar{B}_{i.} - \bar{B}_{..})^2\right\}$$

$$= b(a - 1)\sigma_A^2 + b(a - 1)\sigma_t^2/b.$$

Thus $S_A^2 = b\sum_{i=1}^{a}(\bar{X}_{i.} - \bar{X}_{..})^2/(a - 1) \rightarrow b\sigma_A^2 + \sigma_t^2$.

(*ii*) *Expectation of* S_{BA}^2

$$(X_{ij} - \bar{X}_{i.}) = (B_{ij} - \bar{B}_{i.}),$$

and

$$S_{BA}^2 = \sum^{a}\sum^{b}(X_{ij} - \bar{X}_{i.})^2/a(b - 1) \rightarrow \sigma_t^2 .$$

TABLE 4

Analysis of Variance for Model 1:
Single Treatment Classification Design with
$\underline{b}$ Randomly Sampled Subjects for Each of $\underline{a}$ Levels (True Scores)

Number	Source	Sum of Squares	df	S^2	Expectation of S^2
1	A	$b \sum^{a} (\bar{X}_{i.} - \bar{X}_{..})^2$	$(a-1)$	S_A^2	$b\sigma_A{}^2 + \sigma_t{}^2$
2	B within A	$\sum^{a} \sum^{b} (X_{ij} - \bar{X}_{i.})^2$	$a(b-1)$	$S_{B_A}^2$	$\sigma_t{}^2$
3	Total	$\sum^{a} \sum^{b} (X_{ij} - \bar{X}_{..})^2$	$(ab-1)$		

These outcomes for the analysis of variance are summarized in Table 4. On the null hypothesis $\sigma_A^2 = 0$. One rejects the null hypothesis if the ratio $F_1 = S_A^2/S_{B_A}^2$ with $\text{df}_1 = (a - 1)$ and $\text{df}_2 = a(b - 1)$ exceeds F_α, the tabled value for the chosen level of significance.

Case 3

It is common practice in psychological experimentation to use a design superficially similar to the one just described. That is, one has a series of random samples of subjects allocated to treatment levels and for each subject one has a single score. If, as is usually the case, the scores are fallible, then Model 1 is inapplicable and instead one must write the model to include error of measurement. Assuming that the scores have been drawn from a Case 3 universe, there will be two designs according as one has or has not random replication of measurement on a given subject. For common practice, which provides no measurement replication, Model 3*a* is

$$X_{ij} = A_i + B_{ij} + \Gamma_{ij} .$$

A_i and B_{ij} have been defined above; Γ_{ij} is the random error of measurement component, normally distributed with zero mean and variance σ_e^2. The summary of the analysis of variance for Model 3*a* is given in Table 5. For the test of significance, the null hypothesis is $\sigma_A^2 = 0$. One rejects the null hypothesis if the ratio $F_{3a} = S_A^2/S_{B_A}^2$ with $\text{df}_1 = (a - 1)$ and $\text{df}_2 = a(b - 1)$ exceeds the tabled value of F for the chosen level of significance.

One may note that the terms σ_A^2 and σ_t^2 are common to the expectations of S_A^2 for Models 1 and 3*a*. In addition, the df_1 and df_2 are the same for F_1 and F_{3a}. This enables comparison of the sensitivity of the two tests. The power of the F_1 test is Prob $\{F_1 > F_\alpha \sigma_t^2/(b\sigma_A^2 + \sigma_t^2)\}$; and the power of F_{3a} is Prob $\{F_{3a} > F_\alpha(\sigma_t^2 + \sigma_e^2)/(b\sigma_A^2 + \sigma_t^2 + \sigma_e^2)\}$. The smaller the value to the right of $>$, the greater the power of the test. As $\sigma_t^2/(b\sigma_A^2 + \sigma_t^2) < (\sigma_t^2 + \sigma_e^2)/$

TABLE 5

Analysis of Variance for Model 3a:
Single Treatment Classification Design with
b Randomly Sampled Subjects for Each of *a* Levels (Fallible Scores)

Number	Source	Sum of Squares	df	S^2	Expectation of S^2
1	A	$b\sum^{a}(\bar{X}_{i.}-\bar{X}_{..})^2$	$(a-1)$	S_A^2	$b\sigma_A^2+\sigma_t^2+\sigma_e^2$
2	B within A	$\sum^{a}\sum^{b}(X_{ij}-\bar{X}_{i.})^2$	$a(b-1)$	$S_{B_A}^2$	$\sigma_t^2+\sigma_e^2$
3	Total	$\sum^{a}\sum^{b}(X_{ij}-\bar{X}_{..})^2$	$(ab-1)$		

$(b\sigma_A^2 + \sigma_t^2 + \sigma_e^2)$, the power of F_1 is greater than the power of F_{3a}. That is, analysis in accordance with Model 3*a* provides a less sensitive test of the hypothesis $\sigma_A^2 > 0$ than does Model 1; the loss of sensitivity is due to the intrusion of random error of measurement.

Model 3*a* allows for the acknowledgement of the presence of error variance, but there is no provision for its isolation. To achieve this, one has to add random replication of measurement for each subject. That is, instead of a single score for each subject one has a number of scores. This introduces a source of variation in addition to those already accounted for; accordingly the notation and plan presented above have to be expanded. Let $k = 1, 2, \cdots, c$ represent any one measure or score in a sample of scores C. Then X_{ijk} is the kth score of subject j at treatment level i. As measures on subjects are randomly sampled, k represents number only, not rank. Now Model 3*b* may be written as

$$X_{ijk} = A_i + B_{ij} + \Gamma_{ijk} .$$

A_i and B_{ij} have been defined above; and Γ_{ijk} is defined as was Γ_{ij}. That is, Model 3*a* is the special case of Model 3*b* in which $k = 1$. The summary of the analysis of variance for Model 3*b* is given in Table 6. This analysis provides two tests of significance.

For the first, the null hypothesis is $\sigma_t^2 = 0$. One rejects the null hypothesis if the ratio $F_{3b} = S_{B_A}^2/S_{C_B}^2$ with $\text{df}_1 = a(b-1)$ and $\text{df}_2 = ab(c-1)$ exceeds the tabled value of F for the chosen level of significance. If the null hypothesis is not rejected, the outcome is consistent with the homogeneity of experimental subjects, and in that sense one has zero reliability of measurement. If the null hypothesis is rejected, an estimate of the reliability of measurement may be obtained. With the Case 3 universe, the population value of the reliability coefficient [1] is $\rho_{xx} = \sigma_t^2/(\sigma_t^2 + \sigma_e^2)$, which may be estimated by

TABLE 6

Analysis of Variance for Model 3b:
Single Treatment Classification Design with
$\underline{c}$ Random Measures on each of
$\underline{b}$ Randomly Sampled Subjects for each of
$\underline{a}$ Levels (Fallible Scores)

Number	Source	Sum of Squares	df	S^2	Expectation of S^2
1	A	$bc \sum^{a} (\bar{X}_{i..} - \bar{X}_{...})^2$	(a-1)	S_A^2	$bc\sigma_A{}^2 + c\sigma_t{}^2 + \sigma_e{}^2$
2	B within A	$c \sum^{a} \sum^{b} (\bar{X}_{ij.} - \bar{X}_{i..})^2$	a(b-1)	$S_{B_A}^2$	$c\sigma_t{}^2 + \sigma_e{}^2$
3	C within B	$\sum^{a} \sum^{b} \sum^{c} (X_{ijk} - \bar{X}_{ij.})^2$	ab(c-1)	$S_{C_B}^2$	$\sigma_e{}^2$
4	Total	$\sum^{a} \sum^{b} \sum^{c} (X_{ijk} - \bar{X}_{...})^2$	(abc-1)		

$$r_{zz} = (S^2_{BA} - S^2_{CB})/[S^2_{BA} - S^2_{CB}(1 - c)].$$

For the second, the null hypothesis is $\sigma_A^2 = 0$. One rejects the null hypothesis if the ratio $F'_{3b} = S_A^2/S^2_{BA}$ with $\mathrm{df}_1 = (a - 1)$ and $\mathrm{df}_2 = a(b - 1)$ exceeds the tabled value of F for the chosen level of significance.

Comparison of the power of the F'_{3b} test

$$\text{Prob } \{F'_{3b} > F_\alpha(c\sigma_t^2 + \sigma_e^2)/(bc\sigma_A^2 + c\sigma_t^2 + \sigma_e^2)\}$$

with the powers of F_1 and F_{3a} shows that as

$$\frac{\sigma_t^2}{b\sigma_A^2 + \sigma_t^2} < \frac{c\sigma_t^2 + \sigma_e^2}{bc\sigma_A^2 + c\sigma_t^2 + \sigma_e^2} < \frac{\sigma_t^2 + \sigma_e^2}{b\sigma_A^2 + \sigma_t^2 + \sigma_e^2}$$

then power $F_1 >$ power $F'_{3b} >$ power F_{3a}.

While analysis by the Model 3*b* allows for isolation of an estimate of σ_e^2 , it is important to note that one may *not* convert F'_{3b} to F_1 by subtracting $S^2_{CB} \rightarrow \sigma_e^2$ from the numerator and denominator of F'_{3b} and making appropriate adjustments for the weights b and c. F is the ratio of two independent χ^2 variates—the independence is negated by such a procedure. The only way to achieve the standard of sensitivity of the F_1 test with the given number of subjects is to use error-free measurement. As this is an ideal towards which one can do no more than strive, one has to be satisfied with a less sensitive test. Of the two remaining experimental designs, assuming that one can achieve measurement replication, that which provides the 3*b* form of analysis is to be recommended for general practice. It yields estimates of measurement error variance and reliability, for the latter a test of significance, as well as providing a more sensitive test of treatment effects than the 3*a* design using the same number of subjects. These contentions apply with equal force to the design situations where the t test is ordinarily applied. Finally, while the argument has been in terms of the single treatment classification design, it may be generalized to multiple classification designs.

REFERENCES

[1] Alexander, H. W. The estimation of reliability when several trials are available. *Psychometrika*, 1947, 12, 79-99.

[2] Ehrenberg, A. S. C. The unbiased estimation of heterogeneous error variances. *Biometrika*, 1950, **37**, 347-357.

[3] Walker, Helen M. and Lev, J. *Statistical inference.* New York: Holt, 1953.

Manuscript received 1/14/57
Revised manuscript received 4/30/57

PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

DETERMINATION OF PARAMETERS OF A FUNCTIONAL RELATION BY FACTOR ANALYSIS*

LEDYARD R TUCKER

PRINCETON UNIVERSITY

AND

EDUCATIONAL TESTING SERVICE

Consideration is given to determination of parameters of a functional relation between two variables by the means of factor analysis techniques. If the function can be separated into a sum of products of functions of the individual parameters and corresponding functions of the independent variable, particular values of the functions of the parameters and of the functions of the independent variables might be found by factor analysis. Otherwise approximate solutions may be determined. These solutions may represent important results from experimental investigations.

The possible use of factor analysis techniques to determine parameters of nonlinear functional relations has been a topic for occasional informal discussion. If a factorial approach could be developed it would have considerable application to experimental problems such as learning curves, work decrement curves, dark adaptation curves, etc. This note gives a theoretical basis for determination of parameters by factor analysis for many nonlinear functions.

Factor analytic methods have been limited to investigations applying linear functions of the form (see [2], equation 3, p. 71):

$$s_{ji} = \sum_{m=1}^{r} a_{jm} s_{mi} , \tag{1}$$

where the s_{ji} are the observations, and a_{jm} and s_{mi} are to be estimated. The a_{jm} are task parameters, and the s_{mi} are individual parameters.

In the present context we will consider the functional relation between two variables x and y. Variable x might be termed the independent variable and y might be termed the dependent variable. A general statement of this functional relation for any given individual i is given by

$$y_i = \phi(p_{gi} , x), \tag{2}$$

for which there are a number of parameters p_g which have specific values

*This research was jointly supported in part by Princeton University, the Office of Naval Research under contract N6onr-270-20, and the National Science Foundation under grant NSF G-642.

p_{gi} for each individual. Such a relation is shown graphically in Fig. 1. There exists a family of functions of the form of any given ϕ with the values of p_{gi} defining the particular member of the family. Let j be a particular point of this function with coordinates x_j and y_{ji} . Then

$$y_{ji} = \phi(p_{gi} , x_j). \tag{3}$$

Many functions may be transformed so as to produce

$$y_{ji} = \sum_{m=1}^{r} f_m(x_j)F_m(p_{gi}). \tag{4}$$

The $f_m(x_j)$ are a number of functions of the independent variable x_j . The $F_m(p_{gi})$ are corresponding functions of the parameters p_{gi} . The number, r, of such functions may be finite, or it may be infinite. In this latter case, (4) represents an infinite series, such as Maclaurin's or Taylor's power series or Fourier's trigonometric series (see a standard advanced calculus text, e.g., [1], [3]). Frequently, in this case, a small number of terms of the series will yield an adequate approximation to the y_{ji} . In order to make (1) applicable it is only necessary to define

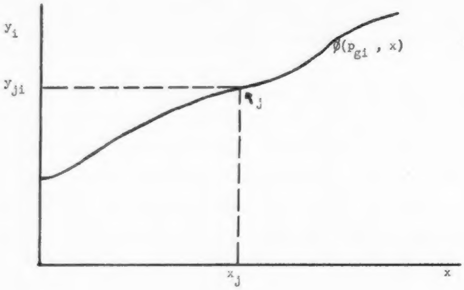


FIGURE 1
A Functional Relation of the Form of (2)

$$a_{jm} \equiv f_m(x_j), \tag{5}$$

$$s_{mi} \equiv F_m(p_{gi}). \tag{6}$$

Then

$$y_{ji} = \sum_{m=1}^{r} a_{jm}s_{mi} . \tag{7}$$

In the present context the s_{mi} will be considered as derived parameters of the transformed function. While they may be expressible in terms of more primitive parameters, they do have the property of determining the particular function for each individual. The family of functions is defined by the a_{jm} . As a consequence of (7), observations of y_{ji} for several given x_j and individuals i may be entered into a score matrix. Each x_j might be used to produce one statistical variable. Estimates of the a_{jm} and s_{mi} then can be obtained by factor analysis techniques.

In order to illustrate the foregoing, consider a learning task for which the learning curve is a simple exponential function, such as

$$y_{ji} = e^{(t_j+b_i)}, \tag{8}$$

where y_{ji} is the performance of individual i on trial j, b_i is a parameter for individual i, and t_j is the number of trials j. t_j replaces x_j as the independent variable in this context, and b_i replaces the parameters p_{gi} . Equation (8) may be transformed to

$$y_{ji} = (e^{t_j})(e^{b_i}). \tag{9}$$

Then

$$a_{j1} = f_1(t_j) = e^{t_j}, \tag{10}$$

$$s_{1i} = F_1(b_i) = e^{b_i}. \tag{11}$$

In this case only one term of the sum of products indicated in (4) and (7) exists. From (9), (10), and (11)

$$y_{ji} = a_{j1}s_{1i} \; . \tag{12}$$

For this simple case, observations are made of the performances on the learning task for each of a number of individuals at each of a selected number of trials. These observations yield a matrix of y_{ji} . A factor analysis will involve a single factor and yield estimates of the a_{j1} and s_{1i} .

The factor analysis problems of communalities and rotation of axes remain to be discussed. In the present context it seems appropriate to assume that each observed y_{ji} may be in error, but the assumption of specific factors seems inappropriate. As a consequence, reliability estimates should be placed in the diagonals of the matrix of intercorrelations. The rotation of axes problem remains unsolved in the present case. The solution is not unique, and the axes may be rotated. It is doubtful, moreover, that the principle of simple structure is applicable when the factor loadings are the various values of the functions $f_m(x_j)$ for the selected points. Some other principle, at present unknown, is needed to fix the location of the axes.

An alternative interpretation of (7) corresponds to the obverse factor procedures, where people are correlated over a population of measures. A large number of values of x_j are selected, and the y_{ji} are observed for a group of individuals. Each of these individuals can be considered as a variable and correlations of the y_{ji} can be obtained for pairs of individuals. The s_{mi} are now the factor loadings, and the a_{jm} are the factor scores. The communalities and rotation of axes aspects of the analysis are quite similar to the corresponding aspects of the first procedure already discussed. One important difference between the present analysis by persons and the previous alternative stems from the more direct determination of the s_{mi} . An inspection of the matrix of s_{mi} might reveal a curvilinear relation between the s_{mi} for several m. Any such relation as the entries in one row being proportional to the square of the entries in another row would indicate a relation to a common, more primitive parameter. The entries in one row being proportional to the product of corresponding entries in two other rows would also be indicative of more primitive parameters. Rotation of axes might be performed so as to reveal such relations.

In any particular situation, the choice as to which variable is to be the independent variable x and which variable is to be the dependent variable y may be quite important. In a learning experiment for a list of paired associates, each trial might be an x_j , and the proportion of correct responses be the observed y_{ji} . However, selected proportions of correct responses might be taken as the x_j , and the numbers of trials necessary to reach these proportions taken as the y_{ji} . Consider a slightly more complex exponential learning curve than that given in (8), such that

$$P = e^{(c_i t + b_i)}, \tag{13}$$

where P is the measure of performance. The parameter c_i has been included as a multiplier to t. This function does not separate in the manner that (8) did unless an infinite series is used. In which case, if values of t_j are chosen and values of P_{ji} are observed, the factor analysis will not involve a definite number of factors. Each successive factor will permit a closer approximation of the series to the function. Some finite number of factors might be found to be adequate.

If logarithms are taken of both sides of (13), it is possible to solve for t as a function of P:

$$t = \frac{1}{c_i} \log P + \frac{b_i}{c_i}. \tag{14}$$

When values of P are selected as P_j and the corresponding t_{ji} are observed, then

$$t_{ji} = \frac{1}{c_i} \log P_j + \frac{b_i}{c_i}. \tag{15}$$

Define

$$a_{j1} \equiv \log P_j , \tag{16}$$

$$s_{1i} \equiv 1/c_i , \tag{17}$$

$$a_{j2} \equiv 1, \tag{18}$$

$$s_{2i} \equiv b_i/c_i . \tag{19}$$

Then

$$t_{ji} = a_{j1}s_{1i} + a_{j2}s_{2i} \ , \tag{20}$$

which is in the form of (7). Only two factors are involved.

Another extension from (8) is to introduce an additive constant d_i:

$$P = d_i + e^{(t+b_i)}. \tag{21}$$

Individual parameters and the variable t may be separated for (21) in the same manner as given for (8). There are now two factors.

If both of the foregoing extensions of (8) are incorporated into a single extension, then

$$P = d_i + e^{(c_i t+b_i)}. \tag{22}$$

The individual parameters do not readily separate now from either variable without employing an infinite series.

It is to be noted that (8) might be treated in the same manner as was (13). The individual parameters might be separated from the variable y or P rather than from t as given. Thus, the foregoing examples include (*i*) a function, equation (8), that may be treated either way; (*ii*) two functions, (13) and (21), each of which may be treated in only one manner; and (*iii*) a function, (22), that cannot be separated. The two single treatment functions form a contrast as to which variable, P or t, is taken as the independent variable. In (13), P should be taken as the independent variable while in (21) t should be taken as the independent variable. In any particular experimental case, the decision as to which variable is to be treated as the independent variable must rest on experience and the judgment of the experimenter. There are cases where the number of factors is excessive whichever variable is taken as the independent variable. The factorial approach may yield in some of these cases an adequate approximation to the observations with a limited number of factors.

REFERENCES

[1] Osgood, W. F. *Advanced calculus.* New York: Macmillan, 1925.
[2] Thurstone, L. L. *Multiple-factor analysis.* Chicago: Univ. Chicago Press, 1947.
[3] Wilson, E. B. *Advanced calculus.* Boston: Ginn, 1911.

Manuscript received 8/15/55

PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

THE INCLUSION OF RESPONSE TIMES WITHIN A STOCHASTIC DESCRIPTION OF THE LEARNING BEHAVIOR OF INDIVIDUAL SUBJECTS

R. J. AUDLEY

UNIVERSITY COLLEGE, LONDON

A stochastic process applicable to the learning behavior of an individual subject is discussed. The process describes both the response times and the sequence of choices obtained from a situation involving two alternatives. Parameter estimates and techniques for assessing goodness of fit are considered.

In a previous paper [2], the possibility of providing a probabilistic description of the learning behavior of an individual subject was discussed. A family of stochastic processes suitable for this purpose was introduced, and problems of parameter estimation and goodness of fit were examined. This examination was restricted to the description of the sequence of responses made by a subject in an experimental situation involving a choice between two alternatives, e.g., the learning of a position habit in a single-unit T-maze. Usually, however, an investigator observes not only the choice made at each trial but also the time taken to make the choice, which for brevity will be referred to here as the response time. The present paper is an attempt to include the response times within the stochastic description elaborated in the earlier paper.

This inclusion of response times carries with it several advantages. The estimation of parameter values can now be based upon a continuous time variable as well as the two-valued variable, success or failure, which was the only datum previously employed. Furthermore, there are certain sequences of responses, such as a long unbroken series of successes or failures, which make it impossible to provide parameter estimates unless response times can be used for this purpose.

The Stochastic Processes

Originally, the processes were based on an urn scheme. Here, however, they will be developed from some simple assumptions, which can be regarded as an identification of the elements of the urn scheme. To give a brief recapitulation of the scheme of the earlier paper: consider an urn containing red and black balls, drawing a red ball being considered equivalent to the occurrence of a correct response, and a black ball to an incorrect response. The number of balls of the two colors is changed after a ball is drawn, according to certain rules. In the present paper, the number of balls of a particular color is identified with a hypothetical mean rate of making the response associated with this color.

For the purpose of simple exposition, attention again will be restricted to data obtained from learning situations involving only two alternative responses, with one response consistently rewarded. At the tth trial, it is assumed that the probability of a correct response occurring in a small time interval $(T, T + \Delta T)$ is $r_t \Delta T$, and of an incorrect response in the same time interval is $w_t \Delta T$. r_t and w_t may be regarded as hypothetical mean rates of responding, i.e., the distribution of response times for either response, *taken individually*, is exponential. This assumption was considered for situations with only one available response by Mueller [10]. Christie [6] has also considered the two-choice situation as one involving the competition between two responses emitted at independent random rates. His paper should be consulted for a more detailed statement of the events supposed to take place at any particular experimental trial.

The probability of no response occurring in time T will be

$$P_0(T) = e^{-(r_t + w_t)T} \tag{1}$$

(e.g., see Feller [7], p. 366).

In the learning situation being considered, the first response to occur terminates an experimental trial. Hence, the probability of a correct response occurring at any trial is the probability that this response is the first to occur. The probability that a correct response terminates the tth trial at time T is from (1) and the basic assumptions equal to

$$e^{-(r_t + w_t)T} r_t \Delta T,$$

and therefore the probability of a success at the tth trial, is

$$P(t) = \int_0^\infty e^{-(r_t + w_t)T} r_t \, dT = \frac{r_t}{r_t + w_t}. \tag{2}$$

It is further assumed that the hypothetical response rates, r_t and w_t, are linear functions of the number of correct and incorrect responses in the first $t - 1$ trials. Thus it is assumed

$$\begin{aligned} r_t &= r_1 + k_t a + (t - 1 - k_t) b, \\ w_t &= w_1 + k_t c + (t - 1 - k_t) d, \end{aligned} \tag{3}$$

where r_1 and w_1 are the initial rates of making correct and incorrect responses, respectively, k_t is the number of correct rewarded responses in the first $(t - 1)$ trials, and a, b, c, and d are parameters associated with the influence of punishment and reward upon the hypothetical response rates.

Substituting for r_t and w_t in (2), the probability of a correct response on the tth trial, given k_t previous successes, is

$$(4) \qquad P(t|k_t) = \frac{r_1 + k_t(a - b) + (t - 1)b}{r_1 + w_1 + k_t(a + c - b + d) + (t - 1)(d + b)}.$$

Dividing numerator and denominator by $(r_1 + w_1)$, and putting

$$\frac{r_1}{r_1 + w_1} = \rho, \qquad \frac{a}{r_1 + w_1} = \alpha, \qquad \frac{b}{r_1 + w_1} = \beta,$$

$$\frac{a + c}{r_1 + w_1} = \gamma_1 , \quad \text{and} \quad \frac{b + d}{r_1 + w_1} = \gamma_2$$

gives

$$(5) \qquad P(t|k_t) = \frac{\rho + k_t(\alpha - \beta) + (t - 1)\beta}{1 + k_t(\gamma_1 - \gamma_2) + (t - 1)\gamma_2}.$$

Equation (5) is the fundamental expression of the earlier paper [2].

The distribution of response times at the tth trial is also completely specified and is exponential. In particular, the mean response time, $\bar{L}_t$, is given by

$$(6) \qquad \bar{L}_t = \int_0^\infty e^{-(r_t + w_t)T}(r_t + w_t)T \, dT = \frac{1}{r_t + w_t}.$$

The relation between response times and probabilities here is based upon the very simplest of assumptions. Clearly, the assumptions concerning the hypothetical response rates and the relation between these rates and past experience can be readily modified. Also, in practice, it is unlikely that the general process, having six parameters, r_1 , w_1 , a, b, c, and d, would be used. Special cases, with some of the parameters a, b, c, and d eliminated, or given particular values, would be more commonly employed. An application of such a special case to experimental data has been given elsewhere [1].

The relation between these stochastic processes and those suggested by other investigators, in particular by Bush and Mosteller [4, 5] and Gulliksen [8, 9], has been fully discussed in the previous paper [2]. However, one further comparison is suggested by the present inquiry. Assumption (3), giving the hypothetical response rates as linear functions of the previous number of correct responses, can be shown to be equivalent to a system of linear operators and is similar to the treatment of a situation with only one available response given by Bush and Mosteller [3]. Thus expression (5) can be included within a linear operator system if the operators are assumed to act not upon the probability of a response but upon response rates hypothetically underlying this probability.

Estimation of Parameters

For brevity of exposition, consider the estimation of parameters for the special case arising when $b = c = 0$ in (3), or equivalently $\alpha = \gamma_1$, $\beta = 0$ in (5). Thus, it is assumed that the effects of reward and punishment of a response are confined to the response rate associated with this particular response and do not generalize to the other. This is the stochastic equivalent of the equation of the learning curve developed by Gulliksen [8, 9].

Consider the data obtained from a simple situation involving a choice between two alternatives. We observe the sequence of choices made by a subject as well as the response time for each trial. The response occurring on the tth trial can be symbolized by a characteristic random variable, X_t. If a correct response occurs, $X_t = 1$; if an incorrect response occurs, $X_t = 0$. Similarly let T_t be the response time at the tth trial. It should be borne in mind, however, that the distribution of possible response times at each trial is taken to be exponential and hence response times close to zero are considered likely. Therefore T_t should more properly be the difference between the response time observed and the minimum response time found in the experimental situation.

Suppose, then, that we have the results of n learning trials of an individual subject, X_t and T_t ($t = 1, 2, \cdots, n$). At the tth trial, the probability of a correct response at time T_t is

$$e^{-(r_t+w_t)T_t} r_t \Delta T,$$

and the probability of an incorrect response at the same time is

$$e^{-(r_t+w_t)T_t} w_t \Delta T.$$

Hence the likelihood, L_n, of the entire sequence of responses and response times is

$$L_n = \prod_{t=1}^{n} [e^{-(r_t+w_t)T_t} r_t^{X_t} w_t^{(1-X_t)}]. \tag{7}$$

We now seek those values of the parameters r_1, w_1, a, and d which maximize L_n. It is more convenient to maximize

$$\lambda_n = \log L_n = \sum_{t=1}^{n} [-(r_t + w_t)T_t + X_t \log r_t + (1 - X_t) \log w_t]. \tag{8}$$

Remembering that $b = c = 0$ is assumed, substitute for r_t and w_t from (3), so that

$$\begin{aligned}\lambda_n = \sum [-(r_1 + w_1 + k_t a + f_t d)T_2 \\ + X_t \log (r_1 + k_t a) + (1 - X_t) \log (w_1 + f_t d)],\end{aligned} \tag{9}$$

where $f_t = t - 1 - k_t$. Differentiating λ_n with respect to r_1, w_1, a, and d, and setting the differentials equal to zero,

$$\frac{d\lambda_n}{dr_1} = -\sum T_t + \sum \frac{X_t}{r_1 + k_t a} = 0; \tag{10}$$

$$\frac{d\lambda_n}{da} = -\sum T_t k_t + \sum \frac{k_t x_t}{r_1 + k_t a} = 0; \tag{11}$$

$$\frac{d\lambda_n}{dw_1} = -\sum T_t + \sum \frac{(1 - X_t)}{w_1 + ftd} = 0; \tag{12}$$

$$\frac{d\lambda_n}{dd} = -\sum T_t f_t + \sum \frac{(1 - X_t) f_t}{w_1 + f_t d} = 0. \tag{13}$$

r_i and w_1 can readily be eliminated. For example, consider (10) and (11). Equation (11) may be rewritten as

$$\sum k_t T_t = \frac{1}{a} \sum X_t \left(1 - \frac{r_1}{r_1 + k_t a}\right) = \frac{1}{a}\left(\sum X_t - r_1 \sum \frac{X_t}{r_1 + k_t a}\right). \tag{14}$$

It should be noted that $k_t = 0$ on the occasion of the first correct response, and hence the summation in (11) extends over one less trial than that in (10). Thus by appropriate substitution from (10),

$$\sum k_t T_t = \frac{1}{a}\left[k - 1 - r_1\left(\sum T_t - \frac{1}{r_1}\right)\right] = \frac{1}{a}(k - r_1 \sum T_t),$$

where k is the total number of correct responses in the entire n learning trials. Hence

$$r_1 = \frac{k - (a \sum k_t T_t)}{\sum T_t}. \tag{15}$$

Similarly, it may be shown that

$$w_1 = \frac{n - k - (d \sum f_t T_t)}{\sum T_t}. \tag{16}$$

Substituting these values for r_1 and w_1 in (11) and (13) two equations are obtained, each in one unknown, namely

$$\begin{aligned} F(a) &= \sum \left\{\frac{k_t x_t}{[(k - a \sum k_t T_t)/(\sum T_t)] + k_t a}\right\} - \sum k_t T_t = 0; \\ F(b) &= \sum \left\{\frac{f_t(1 - x_t)}{[(n - k - d \sum f_t T_t)/(\sum T_t)] + f_t d}\right\} - \sum f_t T_t = 0. \end{aligned} \tag{17}$$

These equations may appear formidable, but they are not difficult to set up and can readily be solved by a numerical iterative procedure. Generally a Taylor series expansion has been employed (e.g., see Whittaker and

Robinson [11]). Having found the appropriate estimates of a and d, (15) and (16) give estimates of r_1 and w_1. If the alternative parameters for the description of the choice sequence alone, ρ, α, and γ_2, are required, the appropriate substitutions are given by (4) and (5).

Goodness of Fit

The stochastic processes described above are nonstationary, and hence no definitive answer to the problems of testing goodness of fit can be given. There are two kinds of data with which the theoretical description can be compared; each comparison presents rather different problems.

Some consideration of the sequence of choices made by a subject has already been given [2]. It was suggested there that the most appropriate procedure would be to determine the distribution of likelihoods of all the possible sequences of length n, given the estimated parameters, and then to compare the likelihood of the observed sequence with this distribution. Unfortunately, as yet, we have been able to determine this distribution only for the simplest case arising from (5), when $\alpha = \beta, \gamma_1 = \gamma_2 = 0$. Lacking any proper statistical procedure, it would apparently be best to compare visually the observed curve of cumulative successes against trial number with a theoretical curve based upon the computed conditional probabilities of success at each trial. Although this is not very satisfactory, it should give some indication of any gross discrepancies between the theoretical description and the experimental data.

In the case of the response times, some idea of the goodness of fit of the stochastic process can be given in the following way. (I am indebted to Dr. D. E. Barton of the Statistics Department, University College, London, for this suggestion.) Having estimated the parameters, the theoretical mean response time at each trial, $\bar{L}_t$, is given by (6). Since the response times are assumed to be distributed exponentially at each trial, the ratio of the observed response times, T_t, to the theoretical mean time $\bar{L}_t$, (i.e., $R_t = T_t/L_t$) should be itself distributed exponentially. Hence exp $(- R_t)$ should have a rectangular distribution in the region (0, 1). Thus the over-all theoretical distribution of response times can be tested against the observed data. Further, a plot of the transformations R_t against the trial number t should reveal any marked trends away from the stochastic description.

Conclusion

It is apparent that answers to the problems of goodness of fit are not very satisfactory. In spite of this, it is suggested that the general approach presented here has some value for the description of experimental data. The procedures given should be sufficient for the comparison of learning behavior occurring under different experimental conditions. Furthermore, the kinds of assumptions underlying the stochastic description make it possible to introduce assumptions concerning the influence of other variables upon learning behavior. In particular, a consideration of the relation between the hypothetical response rates and conditions of motivation might be of some interest.

The basic assumptions may also be modified easily, without changing the general form of the mathematical development. Other theoretical descriptions of learning behavior, therefore, might be readily put into the form suggested by the present paper so that their formulation and verification could be carried out with greater precision.

REFERENCES

[1] Audley, R. J. A stochastic description of the learning behaviour of an individual subject. *Quart. J. exp. Psychol.*, 1957, **9**, 12-20.

[2] Audley, R. J. and Jonckheere, A. R. Stochastic processes for learning. *Brit. J. statist. Psychol.*, 1956, **9**, 87-94.

[3] Bush, R. R. and Mosteller, F. A mathematical model for simple learning. *Psychol. Rev.*, 1951, **58**, 313-323.

[4] Bush, R. R. and Mosteller, F. A stochastic model with applications to learning. *Ann. math. Statist.*, 1953, **24**, 559-585.

[5] Bush, R. R. and Mosteller, F. *Stochastic models for learning.* New York: Wiley, 1955.

[6] Christie, L. S. The measurement of discriminative behavior. *Psychol. Rev.*, 1952, **59**, 443-452.

[7] Feller, W. *An introduction to probability theory and its applications.* New York: Wiley, 1950.

[8] Gulliksen, H. A rational equation of the learning curve based on Thorndike's laws of effect. *J. gen. Psychol.*, 1934, **11**, 395-434.

[9] Gulliksen, H. A generalization of Thurstone's learning function. *Psychometrika*, 1953, **16**, 297-307.

[10] Mueller, C. G. Theoretical relationships among some measures of conditioning. *Proc. nat. Acad. Sci.*, 1950, **36**, 123-130.

[11] Whittaker, E. T. and Robinson, G. *The calculus of observations.* London: Blackie, 1924.

Manuscript received 11/23/56
Revised manuscript received 4/22/57

PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

DETERMINING THE DEGREE OF INCONSISTENCY IN A SET OF PAIRED COMPARISONS*

HAROLD B. GERARD
BELL TELEPHONE LABORATORIES

AND

HAROLD N. SHAPIRO
NEW YORK UNIVERSITY

Consistency in paired comparison data is defined. Two types of inconsistency which may arise are defined. Computational formulas for these types of inconsistency are derived, and examples illustrating the use of these formulas are presented.

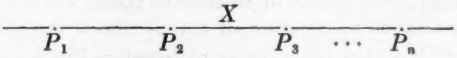
In a recent experiment [1], the authors were concerned with obtaining a measure of *S*'s psychological certainty concerning the probable success of some future undertaking. After exposure to the experimental manipulations, *E* presented *S* with seven 5 × 8 index cards with a different odds for success printed on each card. The stimuli presented were: 10 to 1, 5 to 1, 2 to 1, 1 to 1, 1 to 2, 1 to 5, 1 to 10. All possible pairs of stimuli were presented, and *S* was asked to select the member of each pair which better reflected what he thought his chances were.

From this set of data a measure of both subjective probability of success and *S*'s degree of certainty regarding his estimate was desired. This problem is typical of many in which stimulus comparisons are made. What is presented in this paper is a method for analyzing the consistency of response in such experiments. The method, which involves matrix arithmetic, is quite difficult to formulate in all generality; presented here is a complete analysis of a special case.

The Approach

Let the stimulus cards appear as points P_1, P_2, $\cdots$, P_n on a line which represents a continuum of subjective probability. Let X represent the position of the individual on the line, i.e., his actual subjective probability:

(1)



*These ideas were developed while the first author was on the staff of the Research Center for Human Relations. The work was made possible by the ONR contract NONR 285(10). The authors are indebted to Jack Moshman for his helpful critical suggestions. The United States Government is authorized to reprint this article in whole or in part.

Consider the points, P_i and P_j, and the question, to which of P_i and P_j is X nearer? If P_i is nearer to X than P_j, write $a_{ij} = +1$. If P_j is nearer to X than P_i write $a_{ij} = -1$. Define $a_{ii} = 0$. For all i and j, $a_{ij} = -a_{ji}$.

All of the paired comparisons of the set of points may be tabulated in a square $n \times n$ matrix $A = (a_{ij})$. This matrix has 0 in the principal diagonal. If the row element is closer to X than the column element, the entry is $+1$; contrariwise the entry is -1. Since $a_{ij} = -a_{ji}$ the matrix A is skew symmetric.

The Development

Definition. An *answer matrix* is a skew symmetric $n \times n$ matrix where entries off the main diagonal are all $+1$ or -1.

Definition. The set of responses, or the answer matrix A, is called *inconsistent* if there exists no possible determination of distances between the P_i, and no possible placement of X in (1) for which A is the answer matrix. If some, not necessarily unique, determination of these distances and placement of X is possible then A is called *consistent.*

Definition. For each i, $1 \leq i \leq n$, and a given answer matrix A, define $\lambda_i = \lambda_i(A)$ as the smallest subscript $\lambda_i > i$ such that $a_{i\lambda_i} = +1$. If λ_i does not exist properly define $\lambda_i = \infty$.

Definition. Define $\rho = \rho(A)$ as the *position index* of an answer matrix A as $\rho = \min \lambda_i$. Note that it is possible that $\rho = \infty$, i.e., $\lambda_i = \infty$, for all i, $1 \leq i \leq n$.

THEOREM. *The necessary and sufficient conditions for an answer matrix A to be consistent are that $\rho(A) = \infty$, or that $\rho(A) < \infty$ and there exists a k, $1 \leq k < n$, such that*

$$(i) \quad \rho = \rho(A) = \lambda_k = k + 1,$$

$$(ii) \quad \lambda_k \leq \lambda_{k-1} \leq \cdots \leq \lambda_1 ,$$

$$(iii) \quad \lambda_1 = \begin{cases} i + 1 & \text{for} \quad k \leq i < n \\ \infty & \text{for} \quad i = n, \end{cases}$$

$$(iv) \quad a_{ij} = +1 \quad \text{for} \quad j \geq \lambda_i .$$

These conditions assert that in order to be consistent the skew symmetric matrix A has two connected regions of entries above the principal diagonal, one of $+1$'s and the other of -1's as pictured in Fig. 1. The boundary or demarcation line between the regions appears as "steps" going up and to the right. The case where $\rho = \infty$ is the degenerate case wherein there are no $+1$'s above the diagonal.

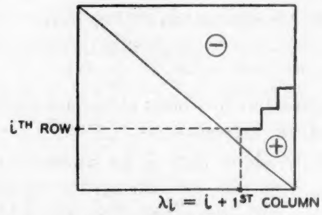


FIGURE 1
Pictorial Representation of the Conditions for Consistency of Matrix A

An examination of the separation diagram (Fig. 1) is in practice the quickest way of deciding whether or not the response matrix is consistent.

PROOF. From the definition of ρ it follows that, when $\rho = \infty$, all the entries above the principal diagonal are -1, and hence those below this diagonal are all $+1$. But this is precisely the answer matrix which corresponds to placing the point X to the right of the last point P_n in (1) or closer to P_n than to P_{n-1}. In the following, then, we may restrict ourselves to the case $\rho < \infty$, i.e., to those cases where X is closer to P_{n-1} than to P_n.

Necessity. Select a k such that X lies between P_k and P_{k+1}. We may assume that X is closer to P_k than to P_{k+1}. (If not, X could be placed between P_{k-1} and P_{k+2} and closer to P_{k+1} without changing any of the answers which determine the entries in the matrix A. This would replace the role of k by $k + 1$, and X would be nearer P_k than P_{k+1}.)

$$\text{(2)} \qquad \underset{P_1}{\cdot} \quad \underset{P_2}{\cdot} \;\cdots\; \underset{P_k}{\cdot} \quad \overset{X}{\cdot} \quad \underset{P_{k+1}}{\cdot} \;\cdots\; \underset{P_n}{\cdot}$$

In (2), $a_{ij} = -1$ for $i < j \le k$, which implies that $\lambda_i > k$ for $i = 1, \cdots, k - 1$. Since $a_{k,k+1} = +1$, $\lambda_k = k + 1$. Also for $i > k$, $j > i$ it is clear that $a_{ij} = +1$, so that $\lambda_i = i + 1$. Thus $\rho = \lambda_k = k + 1$ and conditions (*i*) and (*ii*) are established.

In addition to knowing that X is between P_k and P_{k+1} and closer to P_k, how much additional information is necessary to determine completely the answer matrix A? Suppose, for each i, $1 \le i \le k$, it is known which is the first P_j, $j > i$, such that X is closer to P_i than to P_j. If P_{μ_i} is this point it is clear that

$$a_{ij} = -1 \quad \text{for} \quad j < \mu_i ,$$

and

$$a_{ij} = +1 \quad \text{for} \quad j \ge \mu_i ,$$

and the matrix is completely determined. Clearly also $\lambda_i = \mu_i$ for $1 \le i \le k$.

Since it is immediate from the definition of P_{μ_i} that

$$\mu_k \leq \mu_{k-1} \leq \cdots \leq \mu_1 ,$$

therefore (*ii*) follows. From what has been given above (*iv*) is also immediate. This completes the proof of necessity.

Sufficiency. We wish to show that if an answer matrix A satisfies the conditions (*i*), (*ii*), (*iii*), and (*iv*), we can find a configuration of the distances between the P_i and a position for X in (1) which realizes it. Again assume that $\rho(A) < \infty$. Let $\rho(A) = k$. Place P_k and P_{k+1} on a line with X between them and closer to P_k, as in (2). Consider $\lambda_{k-1} \geq \lambda_k = k + 1$. If $\lambda_{k-1} = k + 1$, place P_{k-1} close to P_k such that

$$\overline{P_{k-1}X} < \overline{XP_{k+1}}$$

(where $\overline{PQ}$ denotes the length of the line segment PQ). If $\lambda_{k-1} > k + 1$, place $P_{\lambda_{k-1}}$ to the right of P_{k+1} and P_{k-1} to the left of P_k such that the following inequalities are satisfied:

$$\overline{P_{k-1}X} > \overline{XP_{k+1}} ,$$
$$\overline{XP_{\lambda_{k-1}}} > \overline{P_{k-1}X}.$$

Next consider $\lambda_{k-2} \geq \lambda_{k-1}$. If $\lambda_{k-2} = \lambda_{k-1}$, place P_{k-2} to the left of, and close to, P_{k-1} such that

$$\overline{P_{k-2}X} < \overline{XP_{\lambda_{k-1}}} .$$

If $\lambda_{k-2} > \lambda_{k-1}$, place P_{k-2} to the left of P_{k-1} and $P_{\lambda_{k-2}}$ to the right of P_{k-1} so that

$$\overline{P_{k-2}X} < \overline{XP_{\lambda_{k-2}}} .$$

If this process stops at a $\lambda_i = \infty$, choose P_i far enough to the left so that

$$\overline{P_iX} > \overline{XP_{\lambda_{i+1}}} ,$$

and place the remaining P_j, $j < i$, to the left of P_i, and the remaining P_h, $h > \lambda_{i+1}$, close to and to the right of $P_{\lambda_{i+1}}$, such that the last point P_h, satisfies

$$\overline{P_iX} > \overline{XP_h} .$$

The resulting configuration clearly has A as its answer matrix. This completes the proof of the sufficiency.

Fundamental types of inconsistency

An answer matrix A may be inconsistent for a variety of reasons. Consider two simple reasons which we designate as fundamental.

Intransitivity. Suppose we have a triplet of subscripts i, j, k such that $a_{ij} = +1$, $a_{jk} = +1$, $a_{ik} = -1$. Then the answer matrix is inconsistent, for if the matrix A is realized by the set of points, P_l , and a position for X, we would have P_i , P_j , P_k as three distinct points with

$$| P_i - X | < | P_j - X | \quad (\text{from } a_{ij} = +1),$$

$$| P_j - X | < | P_k - X | \quad (\text{from } a_{jk} = +1),$$

and

$$| P_k - X | < | P_i - X | \quad (\text{from } a_{ik} = -1).$$

But the first two inequalities imply $| P_i - X | < | P_k - X |$ in contradiction to the third. Thus the matrix A cannot be realized. An inconsistency manifested in this way is called an *intransitivity.*

From the skew symmetry of an answer matrix, the triplet described above, implies also that

$$a_{ki} = +1, \qquad a_{ij} = +1, \qquad a_{kj} = -1,$$

and

$$a_{jk} = +1, \qquad a_{ki} = +1, \qquad a_{ji} = -1.$$

That is, there are apparently three intransitivities generated. In the following we shall count these as *one* intransitivity involving the triplet of subscripts i, j, k.

Separation. Suppose we have a triplet of subscripts $i < j < k$ such that $a_{ij} = + 1$ and $a_{jk} = - 1$; then the answer matrix A is inconsistent. From $a_{ij} = + 1$, X must be to the left of P_j and from $a_{jk} = - 1$, X must be to the right of P_j , which is impossible. An inconsistency manifested by a triplet i, j, k such that $i < j < k$ and $a_{ij} = + 1$, $a_{jk} = - 1$, is called a *separation.*

It is important to note that the cause of an intransitivity is independent of the ordering property of the points, whereas separations are intimately connected with some assumed a priori order requirement. It is not true that intransitivities and separation errors are the only possible errors one can characterize in a set of paired comparisons. It is, however, true that inconsistency as herein defined will result in at least one intransitivity and/or separation error. The concepts of separation and intransitivity are independent in the sense that there exist answer matrices possessing one without the other.

Characterization of consistency

THEOREM. *An answer matrix is consistent if and only if it contains no intransitivities or separations.*

PROOF. The "only if" part of the theorem is quite trivial since the presence of an intransitivity or separation renders an answer matrix inconsistent. On the other hand, assume the answer matrix has no intransitivities or separations. We will prove that it satisfies the conditions (*i*), (*ii*), (*iii*), (*iv*). Clearly we may assume that $\rho(A) < \infty$, since if $\rho(A) = \infty$ the matrix is consistent.

Let the rth column be the first column such that a $+ 1$ appears above the main diagonal; $r > 1$, and it exists since we assume that $\rho(A) < \infty$. We will first prove that above the main diagonal all $- 1$'s in the kth column are above all $+ 1$'s. For if otherwise,

$$a_{ik} = +1, \qquad a_{jk} = -1, \qquad i < j < k,$$

or

$$a_{ik} = +1, \qquad a_{kj} = +1.$$

Since there are no intransitivities this implies $a_{ij} = + 1$. But then $a_{ij} = + 1$, $a_{jk} = - 1$, $i < j < k$ is a separation, which is impossible.

The above argument demonstrates that above the main diagonal $- 1$'s and $+ 1$'s group themselves as required. Now in the rth column (the first column with a $+ 1$ above the main diagonal) we must have $a_{r-1,r} = + 1$, so that $\lambda_{r-1} = r$. Thus there remains to prove only that

$$r = \lambda_{r-1} < \lambda_{r-2} \leq \cdots \leq \lambda_1 , \tag{3}$$

since it then follows that $\rho(A) = r - 1$ and $\lambda_i = i + 1$ for $i \geq r$. Suppose that (3) is false, i.e., for some k, $r - 2 \geq k \geq 1$, $\lambda_k < \lambda_{k+1}$. Then $a_{k\lambda_k} = + 1$, $a_{k+1,\lambda_k} = - 1$ since $\lambda_k \geq r \geq k + 2 > k + 1$. But this contradicts the fact that above the main diagonal $- 1$'s are above $+ 1$'s in each column. We must also verify condition (*iv*) of the consistency theorem, i.e., that for $j \geq \lambda_i$, $a_{ij} = + 1$. Suppose that this is false, i.e., for some $j \geq \lambda_i$, $a_{ij} = - 1$. Since $a_{i\lambda_i} = + 1$, $j > \lambda_i$, then $a_{ji} = + 1$, and since there are no intransitivities $a_{j\lambda_i} = + 1$. It follows that $a_{\lambda_i j} = - 1$ and $a_{i\lambda_i} = + 1$, which is a separation. Thus the conditions of the consistency theorem are satisfied and A is consistent.

Since the notions of intransitivity and separation lie at the basis of the degree to which an answer matrix can be said to be inconsistent, we next consider the question of determining the number of intransitivities and separations.

Number of intransitivities

Let T = the number of intransitivities in an answer matrix A; R_k = the sum of entries in the kth row of A.

THEOREM.

$$T = \frac{1}{24}\left[n(n^2 - 1) - 3 \sum_{k=1}^{n} R_k^2\right].$$

PROOF. For convenience, introduce $C_i = \sum_{j=1}^{n} a_{ji}$ = sum of the entries in the ith column of A. Since $a_{ji} = -a_{ij}$,

$$C_i = -R_i \, . \tag{4}$$

We first consider for a given pair i, j ($i \neq j$) the number of k such that $k \neq i$, j, and

$$a_{ik} = +1, \qquad a_{kj} = +1, \qquad a_{ij} = -1, \tag{5}$$

or

$$a_{jk} = +1, \qquad a_{ki} = +1, \qquad a_{ji} = -1.$$

Let $N_{++}^{(i,j)}$ = no. of k, $1 \leq k \leq n$ such that $a_{ik} = +1$, $a_{kj} = +1$;

$N_{--}^{(i,j)}$ = no. of k, $1 \leq k \leq n$ such that $a_{ik} = -1$, $a_{kj} = -1$;

$N_{+-}^{(i,j)}$ = no. of k, $1 \leq k \leq n$ such that $a_{ik} = +1$, $a_{kj} = -1$;

$N_{-+}^{(i,j)}$ = no. of k, $1 \leq k \leq n$ such that $a_{ik} = -1$, $a_{kj} = +1$.

The superscript (i, j) is omitted in what follows.

Denote by Z the $n \times n$ matrix with zeros on the main diagonal and all other entries $+1$. For any matrix U, $[U]_{i,j}$ denotes the entry in the ith row and jth column of U. Then, for $i \neq j$,

$$(i) \quad n - 2 = N_{++} + N_{--} + N_{+-} + N_{-+} \, ,$$

$$(ii) \quad [AZ]_{i,j} = N_{++} - N_{--} + N_{+-} - N_{-+} \, ,$$

$$(iii) \quad [ZA]_{i,j} = N_{++} - N_{--} - N_{+-} + N_{-+} \, ,$$

and

$$(iv) \quad [A^2]_{i,j} = N_{++} + N_{--} - N_{+-} - N_{-+} \, .$$

Observe in (*ii*) that for N_i^+ = number of k such that $a_{ik} = 1$, and N_i^- = number of k such that $a_{ik} = -1$, we have $[AZ]_{i,j} = N_i^+ - N_i^-$, and $N_i^+ = N_{++} + N_{+-}$, $N_i^- = N_{-+} + N_{--}$. Note in addition that

$$\begin{aligned} [AZ]_{i,j} &= \sum_{k=1}^{n} a_{ik} - a_{ij} = R_i - a_{ij} \, , \\ [ZA]_{i,j} &= \sum_{k=1}^{n} a_{kj} - a_{ij} = C_j - a_{ij} \, , \\ [A^2]_{i,j} &= \sum_{k=1}^{n} a_{ik}a_{kj} \, . \end{aligned} \tag{6}$$

Adding (*ii*) and (*iii*), then (*i*) and (*iv*), one obtains respectively

$$(v) \quad \tfrac{1}{2}([AZ]_{i,j} + [ZA]_{i,j}) = N_{++} - N_{--} \, ,$$

and

$$(vi) \quad \tfrac{1}{2}(n - 2 + [A^2]_{i,j}) = N_{++} + N_{--} \,.$$

Now if $N_{++} \neq 0$ there exists a k such that $a_{ik} = +1$, $a_{kj} = +1$, so that in order to have consistency a_{ij} would have to be $+1$. On the other hand, if $N_{--} \neq 0$ there exists a k such that $a_{ik} = -1$, $a_{kj} = -1$ or $a_{ki} = +1$, $a_{jk} = +1$, and consistency would require that $a_{ji} = +1$ or $a_{ij} = -1$. Therefore if $a_{ij} = +1$, the number of intransitivities involving i, j as in (5) equals N_{--} ; if $a_{ij} = -1$, the number of intransitivities involving i, j as in (5) equals N_{++} . Thus, in any event, the number of intransitivities involving i, j as in (5) equals

$$\tfrac{1}{2}[(1 + a_{ij})N_{--} + (1 - a_{ij})N_{++}] = \tfrac{1}{2}[(N_{--} + N_{++}) - a_{ij}(N_{++} - N_{--})].$$

Using (v), (vi), this in turn equals

$$\tfrac{1}{4}\{(n - 2) + [A^2]_{i,j} - a_{ij}([AZ]_{i,j} + [ZA]_{i,j})\}.$$

From (6) this may be rewritten

$$\mu_{i,j} = \tfrac{1}{4}[(n - 2) + \sum_k a_{ik}a_{kj} - a_{ij}(R_i + C_j - 2a_{ij})].$$

Also,

$$T = \tfrac{1}{6} \sum_{i \neq j} \mu_{i,j} \,.$$

The factor 1/6 arises since in $\mu_{i,j}$, i and j have symmetrical roles so that, in the sum over all unequal i, j, each comparison is counted twice. Also an extra factor of 1/3 is introduced since we do not count as distinct a "permutation" of an intransitivity. Since for $i \neq j$, $a_{kj}^2 = 1$ we may rewrite

$$\mu_{i,j} = \tfrac{1}{4}[n + \sum_k a_{ik}a_{kj} - a_{ij}(R_i + C_j)].$$

Now,

$$\begin{aligned}\sum_{i \neq j} \sum_k a_{ik}a_{kj} &= \sum_{i,j} \sum_k a_{ik}a_{kj} + \sum_i \sum_k a_{ik}^2 \\ &= \sum_k (\sum_i a_{ik})(\sum_j a_{kj}) + n(n - 1) \\ &= \sum_k C_k R_k + n(n - 1) \\ &= -\sum_k R_k^2 + n(n - 1).\end{aligned}$$

Also,

$$\begin{aligned}\sum_{i \neq j} a_{ij}(R_i + C_i) &= \sum_{i,j} a_{ij}(R_i + C_j) \\ &= \sum_i R_i \sum_j a_{ij} + \sum_j C_j \sum_i a_{ij} \\ &= \sum_i (R_i^2 + C_j^2) = 2 \sum_k R_k^2 \,.\end{aligned}$$

Finally, since $\sum_{i \neq j} n = n^2 \, (n - 1)$,

$$
\begin{aligned}
T &= \tfrac{1}{6} \sum_{i \neq j} \mu_{i,j} \\
&= \frac{1}{24} \left(n^2(n - 1) + n(n - 1) - \sum_k R_k^2 - 2 \sum_k R_k^2\right) \\
&= \frac{1}{24} \left[n(n^2 - 1) - 3 \sum_k R_k^2\right],
\end{aligned}
$$

which establishes the theorem. The formula easily may be transformed into a result given by Kendall ([2], p. 156) for what he calls circular triads.

Number of separations

By a method similar to that used above, a formula may be obtained for the number of separations in an answer matrix A. Let $\hat{A}$ = matrix resulting from A by making all entries below the main diagonal equal to zero, and write

$$\hat{A} = (\hat{a}_{ij}),$$

so that $\hat{a}_{ij} = 0$ for $i \geq j$ and $\hat{a}_{ij} = a_{ij}$ for $i < j$. Also let

$\hat{C}_k$ = sum of the entries of the kth column of $\hat{A}$;
$\hat{R}_k$ = sum of the entries of the kth row of $\hat{A}$;
$S = S(A)$ = number of separations in A.

THEOREM.

$$S = \frac{1}{4}\left[\frac{n(n - 1)(n - 2)}{6} + \sum_k \hat{C}_k(n - k) - \sum (k - 1)\hat{R}_k - \sum_k \hat{R}_k\hat{C}_k\right].$$

PROOF. Let $\hat{Z} = (\hat{z}_{ij})$ be the $n \times n$ matrix with $+1$ above the main diagonal and zero elsewhere. We propose to count for a fixed pair, $i < j$, the number of k, $i < k < j$ such that $a_{ik} = +1$ and $a_{kj} = -1$. For $j > i$,

$$
\begin{aligned}
[\hat{A}\hat{Z}]_{i,j} &= \sum_k \hat{a}_{ik}\hat{z}_{kj} \\
&= \sum\nolimits_{\substack{k \\ j>k>i}} a_{ik} \qquad (7) \\
&= \hat{N}_{++}^{(i,j)} - \hat{N}_{--}^{(i,j)} + \hat{N}_{+-}^{(i,j)} - \hat{N}_{-+}^{(i,j)},
\end{aligned}
$$

where $\hat{N}_{++}^{(i,j)}$ = no. of k, $i < k < j$ such that $a_{ik} = +1$, $a_{kj} = +1$ and the others are defined analogously. Where no confusion is possible the superscript (i, j) is omitted.

$$
\begin{aligned}
(8) \qquad [\hat{Z}\hat{A}]_{i,j} &= \sum \hat{z}_{ik}\hat{a}_{kj} = \sum_{j>k>i} a_{kj} \\
&= \hat{N}_{++} - \hat{N}_{--} - \hat{N}_{+-} + \hat{N}_{-+} \, .
\end{aligned}
$$

$$j - 1 - i = \hat{N}_{++} + \hat{N}_{--} + \hat{N}_{+-} + \hat{N}_{-+} \,. \tag{9}$$

$$[A^2]_{i,j} = \sum_{i<k<j} a_{ij} a_{kj} = \hat{N}_{++} + \hat{N}_{--} - \hat{N}_{+-} - \hat{N}_{-+} \,. \tag{10}$$

To solve for $\hat{N}_{+-}^{(i,j)}$,

$$\tfrac{1}{2}\{[\hat{A}\hat{Z}]_{i,j} - [\hat{Z}\hat{A}]_{i,j}\} = \hat{N}_{+-} - \hat{N}_{-+} \,,$$

$$\tfrac{1}{2}\{j - i - 1 - [\hat{A}^2]_{i,j}\} = \hat{N}_{+-} + \hat{N}_{-+} \,,$$

so that

$$\hat{N}_{+-} = \tfrac{1}{4}\{[\hat{A}\hat{Z}]_{i,j} - [\hat{Z}\hat{A}]_{i,j} + j - i - 1 - [\hat{A}^2]_{i,j}\}, \tag{11}$$

and the total number of separations

$$S = \sum_{j>i} \sum \hat{N}_{+-} \,. \tag{12}$$

Note that

$$\begin{aligned}
\sum_{j>i} \Big(\sum_{j>k>i} a_{ik} \Big) &= \sum_{i,j,k} \hat{z}_{ij} \hat{z}_{kj} a_{ik} \\
&= \sum_{j,k} \hat{z}_{kj} \sum_{i<k} a_{ik} \\
&= \sum_{j,k} \hat{z}_{kj} \hat{C}_k \\
&= \sum_{k} \hat{C}_k (n - k).
\end{aligned} \tag{13}$$

$$\begin{aligned}
\sum_{j>i} \Big(\sum_{j>k>i} a_{kj} \Big) &= \sum_{i,j,k} \hat{z}_{ik} \hat{z}_{ij} \hat{a}_{kj} \\
&= \sum_{k,j} \hat{a}_{kj} \sum_{i} \hat{z}_{ik} \hat{z}_{ij} \\
&= \sum_{k,j} \hat{a}_{kj} (k - 1) = \sum_{k} (k - 1) \sum_{j>k} a_{kj} \\
&= \sum_{k} (k - 1)\, R_k \,.
\end{aligned} \tag{14}$$

$$\begin{aligned}
\sum_{j>i} (j - i - 1) &= \sum_{i,j} (j - i - 1) \hat{z}_{ij} \\
&= \sum_{j} (j - 1) \sum_{i} \hat{z}_{ij} - \sum_{i} i \sum_{j} \hat{z}_{ij} \\
&= \sum_{j} (j - 1)^2 - \sum_{i} i(n - i) \\
&= 2 \sum_{j=1}^{n} j^2 - n^2 - \frac{n^2(n + 1)}{2} \\
&= \frac{2n^3 + 3n^2 + n}{3} - \frac{n^3 + 3n^2}{2}
\end{aligned} \tag{15}$$

$$= \frac{n(n-1)(n-2)}{6}.$$

$$\begin{aligned} \sum \hat{a}_{ik}\hat{a}_{kj} &= \sum_{k,i} \hat{a}_{ik} \sum_{j} \hat{a}_{kj} \\ &= \sum_{k,i} \hat{a}_{ik}\hat{R}_k \qquad (16) \\ &= \sum_{k} \hat{R}_k\hat{C}_k\,. \end{aligned}$$

Combining the ten equations, (7) through (16), yields

$$S = \frac{1}{4}\left[\frac{n(n-1)(n-2)}{6} + \sum_k \hat{C}_k(n-k) - \sum_k (k-1)\hat{R}_k - \sum_k \hat{R}_k\hat{C}_k\right].$$

This result completes the proof.

Consideration of the following three examples will clarify the application of the formulas. In all three examples, $n = 7$.

Example I

$$\begin{array}{rrrrrrrrrr}
 & & & & & & & & R_k & \hat{R}_k \\
 & 0 & -1 & -1 & +1 & +1 & +1 & +1 & 2 & 2 \\
 & +1 & 0 & +1 & +1 & +1 & +1 & +1 & 6 & 5 \\
 & +1 & -1 & 0 & +1 & +1 & +1 & +1 & 4 & 4 \\
A = & -1 & -1 & -1 & 0 & +1 & +1 & +1 & 0 & 3 \\
 & -1 & -1 & -1 & -1 & 0 & +1 & +1 & -2 & 2 \\
 & -1 & -1 & -1 & -1 & -1 & 0 & +1 & -4 & 1 \\
 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & -6 & 0 \\
\hat{C}_k & 0 & -1 & 0 & 3 & 4 & 5 & 6 & &
\end{array}$$

$$\tfrac{1}{6}[n(n-1)(n-2)] = 35;$$

$$\sum_k \hat{C}_k(n-k) = -5 + 9 + 8 + 5 = 17;$$

$$\sum_k (k-1)\hat{R}_k = 5 + 8 + 9 + 8 + 5 = 35;$$

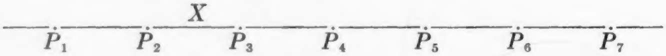
$$\sum_k \hat{R}_k\hat{C}_k = 0 - 5 + 9 + 8 + 5 = 17;$$

$$S = \tfrac{1}{4}(35 + 17 - 35 - 17) = 0;$$

$$\sum R_k^2 = 4 + 36 + 16 + 0 + 4 + 16 + 36 = 112;$$

$$T = \frac{1}{24}[7(48) - 3(112)] = 0.$$

In this matrix there are no separations or intransitivities. The matrix A is consistent. A clear demarcation line exists above the diagonal separating the + 1 and − 1 entries. This boundary appears as steps going up and to the right. The matrix is realized by



This realization is certainly not unique. There are many possible realizations which meet the criteria, namely the set of inequalities which must be satisfied.

Example II

This exemplifies an answer matrix with separations and no intransitivities. Form the matrix A_1 by changing A of *Example I* so that $\alpha_{67} = -1$ and $\alpha_{76} = +1$. The sums then are

R_k	2	6	4	0	−2	−6	−4
$\hat{R}_k$	2	5	4	3	2	−1	0
$\hat{C}_k$	0	−1	0	3	4	5	4

Clearly $T = 0$. Now compute S.

$$\sum_k \hat{C}_k(n - k) = -5 + 0 + 9 + 8 + 5 = 17;$$

$$\sum_k \hat{R}_k(k - 1) = 5 + 8 + 9 + 8 - 5 = 25;$$

$$\sum_k \hat{R}_k\hat{C}_k = 0 - 5 + 0 + 9 + 8 - 5 = 7;$$

$$S = \tfrac{1}{4}(35 + 17 - 25 - 7) = 5.$$

The five separations would in fact be

$$\alpha_{16} = +1, \quad \alpha_{67} = -1;$$

$$\alpha_{26} = +1, \quad \alpha_{67} = -1;$$

$$\alpha_{36} = +1, \quad \alpha_{67} = -1;$$

$$\alpha_{46} = +1, \quad \alpha_{67} = -1;$$

$$\alpha_{56} = +1, \quad \alpha_{67} = -1.$$

It is clear that the difference between the matrix A and the matrix A_1 is that the positions of P_6 and P_7 have been interchanged.

Example III

This exemplifies an answer matrix with intransitivities but no separations.

Consider

$$\begin{array}{r|rrrrrrr|rr}
 & & & & & & & & R_k & \hat{R}_k \\
 & 0 & -1 & -1 & -1 & -1 & -1 & +1 & -4 & -4 \\
 & +1 & 0 & -1 & -1 & -1 & -1 & +1 & -2 & -3 \\
 & +1 & +1 & 0 & -1 & -1 & -1 & +1 & 0 & -2 \\
A = & +1 & +1 & +1 & 0 & -1 & -1 & +1 & 2 & -1 \\
 & +1 & +1 & +1 & +1 & 0 & -1 & +1 & 4 & 0 \\
 & +1 & +1 & +1 & +1 & +1 & 0 & -1 & 4 & -1 \\
 & -1 & -1 & -1 & -1 & -1 & +1 & 0 & -4 & 0 \\
\hat{C}_k & 0 & -1 & -2 & -3 & -4 & -5 & 4 & &
\end{array}$$

$$\sum_k \hat{C}_k(n - k) = -5 - 8 - 9 - 8 - 5 = -35;$$

$$\sum_k (k - 1)\hat{R}_k = -3 - 4 - 3 + 0 - 5 = -15;$$

$$\sum_k \hat{R}_k\hat{C}_k = 3 + 4 + 3 + 5 = 15;$$

$$S = \tfrac{1}{4}(35 - 35 + 15 - 15) = 0.$$

On the other hand,

$$\sum R_k^2 = 16 + 4 + 4 + 16 + 16 + 16 = 72;$$

$$T = \frac{1}{24}(336 - 216) = 5.$$

In general, inconsistent answer matrices will have both separations and intransitivities. As a measure of deviation from consistency the quantity

$$\Phi = \Phi(A) = S + T$$

is suggested. In terms of this measure (since $\Phi = 0$ if and only if $S = T = 0$), a previous theorem provides that A is consistent if and only if $\Phi(A) = 0$.

Summary and Remarks

The problem of determining degree of inconsistency within a set of paired comparisons has been considered. A definition of consistency was given and two fundamental types of inconsistency were defined—namely, intransitivity and separation. The latter is intimately related to an assumed a priori ordering of stimuli. Formulas were given which enable the counting of the number of each type of inconsistency in a set of data. Proofs of these formulas were also provided. It can be shown that separations can occur without intransitivities and vice versa. In general, however, inconsistent data will contain both separations and intransitivities.

The criteria of consistency developed in this paper is made up of two components: intransitivities and separation errors. The counting of separation errors is appropriate only where an a priori ordering is assumed. However, S may violate the assumed order, and re-order the stimuli in a way which appears consistent to him. One may call a set of responses relatively consistent if there exists some ordering relative to which the responses are consistent, i. e., there are no intransitivities. It is in fact easily seen that a set of responses is relatively consistent if and only if there are no intransitivities.

Implicit in the model is that the stimuli are thought of as being presented simultaneously. If one is interested in the effect of order of presentation upon choice and consistency, a modification of the method may be made. One could consider each stimulus as being a composite of the original stimulus with its order of presentation, and treat each composite stimulus as if it were presented simultaneously, i.e., as the stimuli were treated in the above model.

REFERENCES

[1] Gerard, H. B. Some factors affecting an individual's estimate of his probable success in a group situation, *J. abnorm. soc. Psychol.*, 1956, 52, 235-239.

[2] Kendall, M. G. *Rank correlation methods.* (2nd ed.) New York: Hafner, 1955.

Manuscript received 12/10/56
Revised manuscript received 5/13/57

PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

PROPERTIES OF THE ITEM SCORE MATRIX

ANGUS G. MACLEAN

CALIFORNIA TEST BUREAU

A method of deriving from the item score matrix all the usual statistics describing the performance on a test of a group of examinees is given. Since this matrix usually is not actually written out, but is implicit in a set of punched cards, a method of working from a more compact matrix F is described. A numerical example is presented. Applications and advantages of the method are cited, as compared with that of recording only the examinees' test scores and the item difficulties.

Equally Weighted Items

An item score matrix (X) is an N by n rectangular matrix with elements X_{si} all of which are are either 1 or 0. Each row of (X) is a row vector (X_s), which lists the item scores of student s. If items are to be weighted equally the sum of the elements of (X_s) is $\sum_{i=1}^{n} X_{si} = X_s$, the test score of student s. The sum of the test scores of all students in the sample is

$$\sum_{s=1}^{N} X_s = \sum_{s} \sum_{i} X_{si} = T , \tag{1}$$

the sum of all elements of (X).

The column sums of (X) are of interest since

$$\sum_{s=1}^{N} X_{si} = f_i , \tag{2}$$

the number of students responding correctly to item i.

The square of the test score for student s is obtainable by premultiplying the row vector (X_s) by its transpose, a procedure which yields a square symmetric matrix of unit rank:

$$(X_s^2) = (X_s)'(X_s). \tag{3}$$

The sum of all elements of this matrix is X_s^2 .

Some of the operations to be discussed lead to scalar values, others to matrices, the sums of whose elements are those values. For the purposes of clarity, therefore, all symbols for matrices are enclosed in parentheses, while symbols not so enclosed will denote numbers.

The elements of $(X_s)'(X_s)$ are the products $X_{si}X_{sj}$ for student s. Therefore

$$X_s^2 = \sum_{i} \sum_{j} X_{si}X_{sj} . \tag{4}$$

In general, the square of a sum may be obtained by squaring the row vector whose elements are the sum's components, then summing the elements of the square matrix so obtained.

Summing (4) over the N students gives

$$\sum_{s=1}^{N} X_s^2 = \sum_s \sum_i \sum_j X_{si}X_{sj} = S. \tag{5}$$

S is also obtained by summing the elements of a square symmetric matrix (S) obtained by

$$(S) = (X)'(X). \tag{6}$$

It could also be obtained by adding the N matrices (X_s^2) obtained by (3), that is,

$$(S) = (X)'(X) = \sum_{s=1}^{N} (X_s)'(X_s). \tag{7}$$

The side elements of (S) are the cross-product sums S_{ij} of the columns of (X), while the diagonal elements S_i are the result of multiplying the columns by themselves. That is,

$$S_i = \sum_s X_{si}^2 \, , \tag{8}$$

$$S_{ij} = \sum_s X_{si}X_{sj} \, . \tag{9}$$

T and S always denote summation over the N individuals in the sample. They are the statistics used in calculating standard deviations and correlations, as follows:

$$\sigma_i = \frac{\sqrt{L_i}}{N} \, , \tag{10}$$

$$r_{ij} = \frac{L_{ij}}{\sqrt{L_i}\,\sqrt{L_j}} \, , \tag{11}$$

in which

$$L_i = NS_i - T_i^2 \, , \tag{12}$$

$$L_{ij} = NS_{ij} - T_iT_j \, . \tag{13}$$

It so happens, when scores are either 1 or 0, that

$$S_i = T_i = f_i \, , \tag{14}$$

and

$$S_{ij} = f_{ij} \, , \tag{15}$$

where f_i denotes the number of students scoring 1 on item i and f_{ij} the number scoring 1 on both i and j. In other words, counting may be substituted for adding and multiplying; the matrix (S) obtained by the operation $(X)'(X)$ is identical with the F (frequency) matrix described in a recent paper on item selection methods [1]. This matrix (F) can be easily obtained by IBM machines. It should be remarked that (11) yields phi coefficients when scores are dichotomous.

A procedure has thus been given for obtaining the usual descriptive statistics from the matrix of item scores. In addition such a matrix will yield a great deal of other information which a list of test scores will not. From (X) itself the item difficulties (and, of course, their mean and variance) may be obtained as well as the item variances, test scores, and the sum of test scores of those responding correctly to any item. This last statistic is useful in item selection and may be considered as the product of column i with the column of row sums, i.e.,

$$S_{it} = \sum_{s} X_{si}X_{s} . \tag{16}$$

From $(X)'(X)$ we can obtain the same information plus interitem and item-test (point biserial) correlations, Kuder-Richardson reliability estimates, etc. The relevant formulas and item selection procedures are discussed in [1].

Differentially Weighted Items

Consider now the more general case of differentially weighted items. The foregoing discussion and reference deal with the special case in which every item is given a weight of unity in the general formula for a test score composed of a linear sum of weighted item scores:

$$X_{sw} = w_1X_{s1} + w_2X_{s2} + \cdots + w_nX_{sn} . \tag{17}$$

In matrix notation (17) is equivalent to

$$X_{sw} = (X_s)(W)', \tag{18}$$

where (W) is the row vector of item weights:

$$(W) = (w_1, w_2, \cdots, w_n). \tag{19}$$

If a *matrix* of weighted item scores is desired, perform the operation $(X)(D_w)$, where (D_w) is a diagonal matrix with elements w_1 , w_2 , etc. This leaves the rows unsummed, whereas $(X)(W)'$ sums them. The following operations yield the results indicated:

$(X_s)(D_w)$ = row of weighted item scores for student s.

$(X_s)(W)'$ = weighted test score of student s; sum of elements of $(X_s)(D_w)$.

$(X)(D_w)$ = N by n matrix of weighted item scores.

$(X)(W)'$ = column of N weighted test scores; sums of rows of $(X)(D_w)$.

$(D_w)(F)(D_w)$ = square symmetric matrix of order n exhibiting the weighted S_i and S_{ij} values, i.e., sums of squared weighted item scores and sums of their cross products. This is the matrix (S_w).

$(D_w)(F)(W)'$ = column of the n values of S_{it} ; row sums of $(D_w)(F)(D_w)$.

$(W)(F)(W)'$ = S_w, the sum of squared weighted test scores. This is the sum of all elements of $(D_w)(F)(D_w)$.

T_w may be obtained by summing the elements of $(X)(D_w)$ or $(X)(W)'$, and the standard deviation of the weighted test scores will be

$$\sigma_w = \sqrt{NS_w - T_w^2}/N. \tag{20}$$

If the squares of the individual weighted test scores are desired they may be obtained by $(W)(X_s)'(X_s)(W)'$ or by summing the elements of $(D_w)(X_s)'(X_s)(D_w)$, but it would be easier to square individually the elements of $(X)(W)'$ already obtained.

Of course, the column sums of $(X)(D_w)$ are $f_{iw} = w_i f_i$ and the row vector of these is equal to $(f_i)(D_w)$. The weighted S_i and S_{ij} in $(D_w)(F)(D_w)$ are equal to

$$\begin{aligned} S_{iw} &= \sum_s w_i^2 X_{si}^2 \\ &= w_i^2 S_i \ , \end{aligned} \tag{21}$$

and

$$\begin{aligned} S_{ijw} &= \sum_s w_i X_{si} w_j X_{sj} \\ &= w_i w_j S_{ij} \ . \end{aligned} \tag{22}$$

The foregoing techniques are applicable where item analysis is to be performed on a test composed of weighted items. Alternatively, if item scores had been punched 1 or 0 and it was subsequently decided to weight the items differentially, the mean, variance, and reliability of the revised version might be determined by these techniques. The procedure would employ the original F matrix, if F had been determined initially, or would generate F, and then apply the weighting matrices (W) or (D_w) to produce the desired information.

Illustrative Example

Suppose that five students* made the following scores on a set of four

*This N is chosen purely for illustrative convenience. In practice a representative sample of 200 or more cases is recommended to ensure greater reliability of the statistics derived.

items:

$$(X) = \begin{bmatrix}1&0&1&0\\1&0&0&1\\0&1&1&1\\0&1&0&0\\1&1&1&0\end{bmatrix}\begin{array}{c}X_s\\2\\2\\3\\1\\\underline{3}\end{array}$$

$$\begin{array}{lccccl} f_i & 3 & 3 & 3 & 2 & 11 = \sum_s X_s = \sum_i f_i = T. \\ p_i & .60 & .60 & .60 & .40 & \end{array}$$

$$(S) \quad \text{or} \quad (F) = (X)'(X) = \begin{bmatrix}3&1&2&1\\1&3&2&1\\2&2&3&1\\1&1&1&2\end{bmatrix}\begin{array}{c}S_{it}\\7\\7\\8\\\underline{5}\end{array}$$

$$27 = \sum_i S_{it} = \sum_s X_s^2 = S.$$

(This can be checked by squaring the row sums of X.) Usual statistics:

$$\bar{X} = T/N = 2.2. \quad \text{Also, } \bar{X} = \sum_{i=1}^{n} p_i \,.$$

$$\bar{p} = T/nN = .55.$$

$$\sigma^2 = (NS - T^2)/N^2 = 14/25 = .56.$$

$$\mathrm{K\,R}_{20} = \frac{n}{n-1}\left(1 - \frac{N \sum f_i - \sum f_i^2}{NS - T^2}\right) = \frac{4}{3}\left(1 - \frac{55 - 31}{14}\right) = -.95\,.$$

Kuder-Richardson formula 20 is an index of item homogeneity; a negative value indicates a tendency for the items to be negatively intercorrelated. Inspection of (X) confirms this. To obtain the phi coefficient between items 1 and 4,

$$\begin{aligned}\phi_{14} &= \frac{Nf_{14} - f_1 f_4}{\sqrt{Nf_1 - f_1^2}\,\sqrt{Nf_4 - f_4^2}} \\ &= \frac{-1}{\sqrt{6}\,\sqrt{6}} \\ &= -.17.\end{aligned}$$

In many situations (but usually not in item selection) an L matrix is derived from (S), with side elements $L_{ij} = NS_{ij} - T_iT_j$, and diagonal elements $L_i = NS_i - T_i^2$. The side elements are the numerators of the correlation coefficients, the denominators are the geometric means of the appropriate diagonal elements. In matrix notation

$$(L) = N(S) - (T)'(T), \tag{23}$$

where (T) is a row vector containing the sums of scores on each variable and N is, of course, a scalar. In the case of items scored 1 or 0 this becomes

$$(L) = N(F) - (f)'(f), \tag{24}$$

where (f) is the row vector of item frequencies (number of students scoring 1 on each item). Then, in the example,

$$(L) = \begin{bmatrix} 15 & 5 & 10 & 5 \\ 5 & 15 & 10 & 5 \\ 10 & 10 & 15 & 5 \\ 5 & 5 & 5 & 10 \end{bmatrix} - \begin{bmatrix} 9 & 9 & 9 & 6 \\ 9 & 9 & 9 & 6 \\ 9 & 9 & 9 & 6 \\ 6 & 6 & 6 & 4 \end{bmatrix} = \begin{bmatrix} 6 & -4 & 1 & -1 \\ -4 & 6 & 1 & -1 \\ 1 & 1 & 6 & -1 \\ -1 & -1 & -1 & 6 \end{bmatrix}.$$

It is evident from (L) that four out of the six interitem correlations are negative. It may be noted that the L matrix may be converted into an item covariance matrix by dividing every element by N^2.

Now suppose that it is desirable to apply a set of weights to the items as follows:

$$\begin{array}{rcccc} \text{Item Number} & 1 & 2 & 3 & 4 \\ (W) = & (3 & 3 & 5 & 1) \end{array}$$

Then:

$$\begin{array}{ll} & \text{Row sums of } (X)(D_w) \\ & = (X)(W)' = X_{\cdot w} = \sum_i w_i X_{\cdot i} \, . \\ (X)(D_w) = \begin{bmatrix} 3 & 0 & 5 & 0 \\ 3 & 0 & 0 & 1 \\ 0 & 3 & 5 & 1 \\ 0 & 3 & 0 & 0 \\ 3 & 3 & 5 & 0 \end{bmatrix} & \begin{matrix} 8 \\ 4 \\ 9 \\ 3 \\ \underline{11} \\ 35 = T_w \end{matrix} \end{array}$$

and

$$(D_w)(F)(D_w) = \begin{bmatrix} 27 & 9 & 30 & 3 \\ 9 & 27 & 30 & 3 \\ 30 & 30 & 75 & 5 \\ 3 & 3 & 5 & 2 \end{bmatrix} \qquad \begin{array}{l} \text{Row sums} = (D_w)(F)(W)' \\ 69 \\ 69 \\ 140 \\ \underline{13} \\ 291 = (W)(F)(W)' = S_w \, . \end{array}$$

$$\bar{X}_w = T_w/N = 7.0,$$

$$\sigma_w^2 = (NS_w - T_w^2)/N^2 = 230/25 = 9.25.$$

REFERENCE

[1] MacLean, A. G. and Tait, A. T. Some computational short-cuts in the development or analysis of tests. *J. appl. Psychol.*, 1954, **38**, 260-263.

Manuscript received 3/6/57
Revised manuscript received 5/16/57

PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

THE COUNSELING ASSIGNMENT PROBLEM*

JOE H. WARD, JR.

AIR FORCE PERSONNEL AND TRAINING RESEARCH CENTER

A disposition index, DI, which provides information about each possible placement to be considered in a personnel classification situation is discussed. The index is readily computed by machine methods and can be used by counselors required to make assignments. The use of the disposition index provides an adequate approximation to optimal solutions obtained by other methods.

The personnel classification problem has been discussed previously by several authors [1, 2, 4, 5]. This problem has been shown to be similar to the Hitchcock-Koopmans transportation problem, which is a special case of linear programming [6]. The techniques presented in the following discussion have a direct analogy to the problem of a transportation scheduling supervisor who is responsible for transporting products from several origins to several destinations in an economical manner.

The problem of assigning personnel to jobs generally has been stated as follows [6]: Given n persons to be assigned to n jobs and the productivity of the ith person on the jth job, find an assignment of persons to jobs such that total productivity is a maximum. A solution to this problem can be determined by linear programming techniques [2, 3, 6]; if the problem is not too large, the assignments can be determined by automatic methods without the intervention of counselors. This problem is of particular concern in military and large industrial personnel assignments but is not closely related to individual vocational guidance.

A major difficulty with this approach to the problem is that the productivity values are generally only crude estimates of the value of a person on a job. Consequently there is still need for intervention by counselors to account for unforeseen significant information. An additional problem in the use of a completely counselor-free assignment procedure is that it is quite difficult to sell, operationally. This is probably due, in part, to the drastic, noticeable system change brought about by conversion from the old to the completely automated system.

A reasonable approach indicates continuing the present counseling systems and providing increasingly valuable assignment information that

*This report is based on work done under ARDC Project No. 7702, Task No. 17051, in support of the research and development program of the Air Force Personnel and Training Research Center, Lackland Air Force Base, Texas. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government.

will lead to the optimal solution. Continuous gradual improvement of the information supplied to the counselor assignment process will result in more effective assignments. The procedure may ultimately converge to an automatic system—human intervention decreasing with increasing adequacy of productivity information. This procedure will have the advantage of gradual implementation—leading readily to acceptance because of minimum interference with existing procedures, and more adequate utilization of personnel. The following material will include a description of a placement or disposition index which can fit into a counselor assignment system.

A Counseling Assignment Problem

Consider the problem of assigning n men to n jobs given the productivity, c_{ij} , of the ith man on the jth job. In the counseling situation it would be desirable to have information (perhaps represented by a single index) associated with each possible placement that would reflect characteristics of the entire c_{ij} array. In order to consider the relative merits of particular placements, a counselor should have not only an individual assignee's productivities (as indicated by an aptitude score, achievement score, or some other measure) but also an indication of the productivities of all other personnel to be placed.

Assume that an individual counselor is required to assign three men to three jobs, and suppose the productivity index matrix is as follows:

		Jobs		
		1	2	3
Persons	1	8	7	6
	2	5	1	0
	3	6	4	1

Assume further that the counselor can see only one man's productivities (or perhaps test scores) at a time and that he adopts the policy of placing a man in an available job in which he has the highest productivity. If the men come to the counselor in the above order, the assignment would be as follows:

Person	*Job*
1	1
2	2
3	3

The first man's highest index is on job one; the counselor will therefore place man one on job one. There are then two jobs remaining; since man two has a higher productivity on job two than on job three, he will have job two. Finally, the third man will be placed on job three. This sequence was selected as an example because it would provide the lowest possible sum, $c_{11} + c_{22} + c_{33} = 10$, and therefore would be considered the worst assignment. The maximum sum, $c_{21} + c_{32} + c_{13} = 15$, would have resulted only if the men had entered in the sequence 2, 3, 1. If there had been a completely automatic system which would give the optimal assignment, $c_{13} + c_{21} + c_{32} = 15$, all would be well if there were no possibilities of additional information about productivities.

Assume now that the counselor has determined (before talking with the men) the optimal assignment and feels confident of his position. When man one enters, the counselor plans to place him on job three where his productivity is six. However, after further investigation, the counselor finds it is impossible to make this placement; for lack of a second recommended placement, the counselor places the first man on job one (productivity equal to 8) in his effort to maximize the assignment sum. It is now apparent that the counselor is on his way to making the worst placements again and will be forced into the minimum assignment sum $c_{11} + c_{22} + c_{33} = 10$.

Even though this example is made to demonstrate the worst situation, it is still apparent that it would be desirable to provide the counselor with information reflecting the relative merits of each placement. The disposition index, DI, that is to be developed should provide this type of information and should be expected to result in efficient assignments at small computational expense.

Development of a Disposition Index, DI

Consider, first, placing the person p on the job q. Having made that placement, assume that all possible assignments are made and that each assignment of the $n - 1$ persons is equally likely. Then there are $(n - 1)!$ possible sums containing c_{pq} and the probability associated with each is $1/(n - 1)!$.

Now consider the mean value, $E(S_{pq})$, of the assignment sums containing c_{pq}, and consequently the mean value, $E(s_{pq}) = E(S_{pq})/n$ of the productivities contained in the $(n - 1)!$ sums involving c_{pq}. Having selected the value c_{pq}, the sums contain only elements from the $(n - 1)$ remaining rows and columns of the c_{ij} array. Now each element, say c_{rs}, of the resulting square matrix of order $(n - 1)$ is contained in $(n - 2)!$ of the $(n - 1)!$ sums. Therefore it follows that the mean value $E(S_{pq})$, and consequently $E(s_{pq})$, are obtained as follows:

$$E(S_{pq}) = [(n - 1)!c_{pq} + (n - 2)!(c_{..} - c_{p.} - c_{.q} + c_{pq})]/(n - 1)! ,$$

where

$$c_{..} = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} , \quad c_{p.} = \sum_{j=1}^{n} c_{pj} , \quad c_{.q} = \sum_{i=1}^{n} c_{iq} ;$$

$$\begin{aligned} E(S_{pq}) &= [(n-1)c_{pq} + c_{\cdot\cdot} - c_{p\cdot} - c_{\cdot q} + c_{pq}]/(n-1), \\ &= [nc_{pq} - c_{p\cdot} - c_{\cdot q} + c_{\cdot\cdot}]/(n-1), \end{aligned} \tag{1}$$

and, dividing by n,

$$E(s_{pq}) = \frac{1}{n} E(S_{pq}) = \frac{1}{n(n-1)} [nc_{pq} - c_{p\cdot} - c_{\cdot q} + c_{\cdot\cdot}]. \tag{2}$$

Now consider the mean value, $E(S_{\overline{pq}})$, of the sums *not* containing c_{pq} ; consequently, the mean value $E(s_{\overline{pq}}) = E(S_{\overline{pq}})/n$ of the productivities contained in sums *not* involving c_{pq} . There are $(n-1)$ $(n-1)!$ such sums, and the values of $E(S_{\overline{pq}})$ and $E(s_{\overline{pq}})$ are obtained as follows:

$$\begin{aligned} E(S_{\overline{pq}}) &= \frac{1}{(n-1)(n-1)!} [(n-1)!c_{\cdot\cdot} - (n-1)!c_{pq} \\ &\qquad - (n-2)!(c_{\cdot\cdot} - c_{p\cdot} - c_{\cdot q} + c_{pq})] \\ &= \frac{1}{(n-1)^2} [(n-2)c_{\cdot\cdot} + c_{p\cdot} + c_{\cdot q} - nc_{pq}], \end{aligned} \tag{3}$$

and, dividing by n,

$$E(s_{\overline{pq}}) = \frac{1}{n} E(S_{\overline{pq}}) = \frac{1}{n(n-1)^2} [(n-2)c_{\cdot\cdot} + c_{p\cdot} + c_{\cdot q} - nc_{pq}]. \tag{4}$$

Now consider the difference $D_{pq} = E(S_{pq}) - E(S_{\overline{pq}})$, between the mean sum obtained when placing the pth person on the qth job and *not* making that particular placement. From (1) and (3),

$$\begin{aligned} D_{pq} &= \frac{1}{(n-1)} [nc_{pq} - c_{p\cdot} - c_{\cdot q} + c_{\cdot\cdot}] - \frac{1}{(n-1)^2} [(n-2)c_{\cdot\cdot} + c_{p\cdot} \\ &\qquad + c_{\cdot q} - nc_{pq}] \\ &= \frac{1}{(n-1)^2} [n(n-1)c_{pq} + (n-1)c_{\cdot\cdot} - (n-1)(c_{p\cdot} + c_{\cdot q}) \\ &\qquad - (n-2)c_{\cdot\cdot} - (c_{p\cdot} + c_{\cdot q}) + nc_{pq}] \\ &= \frac{1}{(n-1)^2} [n^2 c_{pq} - n(c_{p\cdot} + c_{\cdot q}) + c_{\cdot\cdot}], \end{aligned} \tag{5}$$

and, dividing by n,

$$\begin{aligned} d_{pq} &= E(s_{pq}) - E(s_{\overline{pq}}) = \frac{1}{n} E(S_{pq}) - \frac{1}{n} E(S_{\overline{pq}}) = D_{pq}/n \\ &= \frac{1}{n(n-1)^2} [n^2 c_{pq} - n(c_{p\cdot} + c_{\cdot q}) + c_{\cdot\cdot}]. \end{aligned} \tag{6}$$

The value D_{pq} represents the difference between the mean value of the assignment sums involving c_{pq} and the mean value of the assignment sums *not* involving c_{pq} . The value d_{pq} represents the difference between the mean value of the productivities contained in assignment sums involving c_{pq} and the mean value of the productivities contained in assignment sums *not* involving c_{pq} . It is apparent then that as the value of d_{pq} or D_{pq} increases the placement of person p on job q is more likely to result in a larger assignment sum.

Some interesting properties of these equations are the following:

$$\sum_{i=1}^{n} E(S_{iq}) = \sum_{j=1}^{n} E(S_{pj}) = \sum_{i=1}^{n} E(S_{\overline{iq}}) = \sum_{j=1}^{n} E(S_{\overline{pj}}) = c_{..}\ ; \tag{7}$$

$$\sum_{i=1}^{n} E(s_{iq}) = \sum_{j=1}^{n} E(s_{pj}) = \sum_{i=1}^{n} E(s_{\overline{iq}}) = \sum_{j=1}^{n} E(s_{\overline{pj}}) = c_{..}/n; \tag{8}$$

consequently,

$$\sum_{i=1}^{n} D_{iq} = \sum_{j=1}^{n} D_{pj} = \sum_{i=1}^{n} d_{iq} = \sum_{j=1}^{n} d_{pj} = 0. \tag{9}$$

This indicates that the values of D_{pq} and d_{pq} are in a type of deviation form simultaneously by rows by columns. Putting the c_{ij} matrix in deviation form by rows (or columns) first and then in deviation form by columns (or rows), the deviational form, δ_{pq} , becomes

$$\delta_{pq} = \frac{1}{n^2}\,[n^2 c_{pq} - n(c_{p.} + c_{.q}) + c_{..}]. \tag{10}$$

Therefore it can be seen that δ_{pq} , obtained by putting the c_{ij} matrix in deviation form by rows and columns, differs from D_{pq} only with respect to the factor $1/n^2$, whereas D_{pq} involves the factor $1/(n-1)^2$.

Since it is frequently desired to assign m persons to n jobs, where $m \geqq n$, consider the expression for D_{pq} and d_{pq} under these more general conditions.

$$E(S_{pq}) = \frac{(m-n)!}{(m-1)!}\left[\frac{(m-1)!}{(m-n)!}\,c_{pq} + \frac{(n-2)!}{(m-n)!}\,(c_{..} - c_{p.} - c_{.q} + c_{pq})\right],$$

where

$$c_{..} = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\ , \quad c_{.q} = \sum_{i=1}^{m} c_{iq}\ , \quad c_{p.} = \sum_{j=1}^{n} c_{pj}\ ;$$

$$\begin{aligned} E(S_{pq}) &= \frac{1}{m-1}\,[(m-1)c_{pq} + c_{..} - c_{p.} - c_{.q} + c_{pq}] \\ &= \frac{1}{m-1}\,[mc_{pq} - c_{p.} - c_{.q} + c_{..}], \end{aligned} \tag{11}$$

and

$$E(s_{pq}) = \frac{1}{n}E(S_{pq}) = \frac{1}{n(m-1)}[mc_{pq} - c_{p.} - c_{.q} + c_{..}]; \tag{12}$$

$$\begin{aligned} E(S_{\overline{pq}}) &= \frac{(m-n)!}{(m-1)(m-1)!}\Bigg[\frac{(m-1)!}{(m-n)!}c_{..} - \frac{(m-1)!}{(m-n)!}c_{pq} \\ &\qquad - \frac{(m-2)!}{(m-n)!}(c_{..} - c_{p.} - c_{.q} + c_{pq})\Bigg] \\ &= \frac{1}{(m-1)^2}[(m-2)c_{..} + c_{p.} + c_{.q} - mc_{pq}], \end{aligned} \tag{13}$$

and

$$E(s_{\overline{pq}}) = \frac{1}{n}E(S_{\overline{pq}}) = \frac{1}{n(m-1)^2}[(m-2)c_{..} + c_{p.} + c_{.q} - mc_{pq}]. \tag{14}$$

Then the difference $D_{pq} = E(S_{pq}) - E(S_{\overline{pq}})$ leads to the expression

$$D_{pq} = \frac{1}{(m-1)^2}[m^2c_{pq} - m(c_{p.} + c_{.q}) + c_{..}]. \tag{15}$$

Dividing by n gives

$$d_{pq} = \frac{1}{n}D_{pq} = \frac{1}{n(m-1)^2}[m^2c_{pq} - m(c_{p.} + c_{.q}) + c_{..}]. \tag{16}$$

It is important to notice the similarities to the several expressions previously developed. We can write

$$\begin{aligned} E(S_{pq}) &= k_1[nc_{pq} - c_{p.} - c_{.q}] + k_2 , \\ E(S_{\overline{pq}}) &= k_3[nc_{pq} - c_{p.} - c_{.q}] + k_4 , \\ D_{pq} &= k_5[nc_{pq} - c_{p.} - c_{.q}] + k_6 , \\ d_{pq} &= k_7[nc_{pq} - c_{p.} - c_{.q}] + k_8 . \end{aligned}$$

Thus if the magnitude of any of these indices is used as a basis for assignment, then the value

$$\phi_{pq} = nc_{pq} - c_{p.} - c_{.q} \tag{17}$$

will provide all of the distinguishing information among possible placements. The easily computed index ϕ_{pq} provides a large amount of information concerning the array of productivities.

There are several possible indices from which a disposition index, DI, may be chosen; the one probably most meaningful to the counselor is (2),

$E(s_{pq})$. This index is the mean value of the productivities contained in all possible assignment sums involving c_{pq} . It is directly related to the productivities, and it has the same interpretation for any value of n. For the more general case of m persons assigned to n jobs, where $m \geqq n$, $E(s_{pq})$ is given by (12). It is therefore suggested that the disposition index DI_{pq} be defined by (12):

$$\mathrm{DI}_{pq} = \frac{1}{n(m-1)}\,[mc_{pq} - c_{p.} - c_{.q} - c_{..}], \tag{18}$$

where m = number of persons to be assigned,
n = number of jobs to be filled,
c_{pq} = productivity of the pth person on the qth job,

$$c_{p.} = \sum_{j=1}^{n} c_{pj}\ ,\ c_{.q} = \sum_{i=1}^{m} c_{iq}\ ,\ c_{..} = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\ .$$

The Disposition Index in a Counseling Assignment System

The disposition index, DI, reflects the relative merits of making a particular placement based upon information about the entire productivity array. The first step in using the DI would be to compute the entire matrix of DI_{pq} ; that is, compute DI_{pq} for every person on every job. If the entire DI matrix is available, placement could proceed by placing the largest DI first, next largest second, and so on until all placements have been made. If elaborate data processing equipment is available, the DI matrix can be computed after each placement to reflect the change of conditions. This should tend to provide an assignment sum that is very nearly optimal. In any case, the reduced matrix of DI can be computed after, say, every tth placement with the frequency of updating determined by the speed of available computing facilities. In actual operation, it would probably be desirable to update the DI matrix at the end of each day and at the same time distribute to counselors the DI's of the personnel to be placed the following day.

Consider the application of DI's to the simple problem presented previously. The productivity array, complete with row and column sums, is:

		Jobs			
		1	2	3	$c_{p.}$
	1	8	7	6	21
Persons	2	5	1	0	6
	3	6	4	1	11
	$c_{.q}$	19	12	7	38 = $c_{..}$

It is now possible to compute the DI matrix.

$$\mathrm{DI}_{pq} = \frac{1}{3(2)} [3c_{pq} - c_{p.} - c_{.q} + c_{..}].$$

$$\mathrm{DI}_{11} = \tfrac{1}{6}[3(8) - 21 - 19 + 38] = \tfrac{1}{6}[24 - 21 - 19 + 38]$$
$$= \tfrac{1}{6}[22] = 22/6.$$

$$\mathrm{DI}_{12} = \tfrac{1}{6}[3(7) - 21 - 12 + 38] = \tfrac{1}{6}[21 - 21 - 12 + 38]$$
$$= \tfrac{1}{6}[26] = 26/6.$$

$$\mathrm{DI}_{13} = \tfrac{1}{6}[3(6) - 21 - 7 + 38] = \tfrac{1}{6}[18 - 21 - 7 + 38]$$
$$= \tfrac{1}{6}[28] = 28/6.$$

The complete set of DI's is obtained by similar computations.

		Jobs			
		1	2	3	$\sum_{j=1}^{3} \mathrm{DI}_{pj}$
Persons	1	22/6	26/6	28/6	38/3
	2	28/6	23/6	25/6	38/3
	3	26/6	27/6	23/6	38/3
$\sum_{i=1}^{3} \mathrm{DI}_{iq}$		38/3	38/3	38/3	$38 = c_{..}$

In this problem the three highest DI's can be selected and the indicated placements made. Man one would be placed on job number three, man two on job one, and man three on job two; this would result in the maximum sum $c_{13} + c_{21} + c_{32} = 15$.

Notice what the counselor would do if man one could not, for some valid reason, be placed on job three. The counselor would *not* place man one on job one as indicated by his highest productivity but would place him in job two where his second highest DI is located, DI $= 26/6$. After making this assignment, the counselor would continue to fill the jobs according to values of the disposition index. The result would be an assignment that has the second highest possible value $c_{12} + c_{21} + c_{33} = 13$.

The next example is selected to demonstrate when the procedure will not give a maximum assignment sum if only one DI matrix is computed. Consider the productivity array shown below.

		Jobs			
		1	2	3	$c_{p.}$
	1	1	4	0	5
Persons	2	1	7	6	14
	3	4	7	7	18
	$c_{.q}$	6	18	13	$37 = c_{..}$

The DI matrix is:

		Jobs			
		1	2	3	$\sum_{j=1}^{3} \mathrm{DI}_{pj}$
	1	29/6	26/6	19/6	37/3
Persons	2	20/6	26/6	28/6	37/3
	3	25/6	22/6	27/6	37/3
	$\sum_{i=1}^{3} \mathrm{DI}_{iq}$	37/3	37/3	37/3	$37 = c_{..}$

The three highest values of DI_{pq} are $\mathrm{DI}_{11} = 29/6$, $\mathrm{DI}_{23} = 28/6$, and $\mathrm{DI}_{33} = 27/6$. Since DI_{23} and DI_{33} involve the same job, if man one is placed on job one, and man two is placed on job three, then it will be necessary to place man three on job two. This assignment will result in a sum which is not optimum, $c_{11} + c_{23} + c_{32} = 14$. However, if after placing the first man on job one, a new DI matrix is computed, an optimal sum will result.

		Jobs		
		2	3	$\sum_{j=2}^{3} \mathrm{DI}_{pj}$
Persons	2	14/2	13/2	27/2
	3	13/2	14/2	27/2
	$\sum_{i=2}^{3} \mathrm{DI}_{iq}$	27/2	27/2	$27 = c_{..}$

From the new DI array it is clear that man two should be placed on job two and man three on job three. This would result in the maximum sum $c_{11} + c_{22} + c_{33} = 15$.

Now consider a much larger assignment problem which involves assignment of three different kinds of people to five different kinds of jobs. The following array presents, rather than productivities, values which might represent the cost of having a person type in a particular type job. The matrix is bordered by the frequencies of men available and jobs to be filled, as well as by row and column totals.

				Job Types				
		1	2	3	4	5	Persons Available	$c_{p\cdot}$
Person Types	1	57	60	55	54	62	40	13940
	2	53	52	50	59	51	80	12890
	3	58	63	61	56	64	120	14550
Job Quota		10	20	30	80	100	240	
	$c_{\cdot q}$	13480	14120	13520	13600	14240	3,334,800 = $c_{\cdot\cdot}$	

A solution based upon such a cost matrix requires a minimization rather than a maximization process. Consequently, it will be necessary to select the smallest values of the DI matrix. From the marginal totals it is then possible to compute the DI matrix.

$$\mathrm{DI} = \frac{1}{57{,}630}\begin{bmatrix} 3{,}321{,}060 & 3{,}321{,}140 & 3{,}320{,}540 & 3{,}320{,}220 & 3{,}321{,}500 \\ 3{,}321{,}150 & 3{,}320{,}270 & 3{,}320{,}390 & 3{,}322{,}470 & 3{,}319{,}910 \\ 3{,}320{,}690 & 3{,}321{,}250 & 3{,}321{,}370 & 3{,}320{,}090 & 3{,}321{,}370 \end{bmatrix}$$

Starting with the smallest value and placing the personnel in ascending order of DI, the following minimum sum assignment is obtained:

				Job Types			
		1	2	3	4	5	Persons Available
Person Types	1		10	30			40
	2					80	80
	3	10	10		80	20	120
Job Quota		10	20	30	80	100	240

The sum associated with this assignment is

$$10(60) + 30(55) + 80(51) + 10(58) + 10(63) + 80(56) + 20(64) = 13{,}300.$$

This example provides an optimal sum without recomputing the DI matrix.

Other Possible Disposition Indexes

It is possible to consider the variances associated with the expected sums and obtain more information about the distribution of sums associated with each possible placement decision. The variances can be easily computed by machine methods and might be incorporated into a useful disposition index.

REFERENCES

[1] Brogden, H. E. An approach to the problem of differential prediction. *Psychometrika*, 1946, **11**, 139-154.

[2] Dwyer, P. S. Solution of the personnel classification problem with the method of optimal regions. *Psychometrika*, 1954, **19**, 11-26.

[3] Dwyer, P. S. The detailed method of optimal regions. *Psychometrika*, 1957, **22**, 43-52.

[4] Thorndike, R. L. The problem of classification of personnel. *Psychometrika*, 1950, **15**, 215-235.

[5] Votaw, D. F., Jr. Methods of solving some personnel-classification problems. *Psychometrika*, 1952, **17**, 255-266.

[6] Votaw, D. F., Jr. and Dailey, J. T. Assignment of personnel to jobs. Research Bulletin 52-24, Air Training Command, Human Resources Research Center. Lackland Air Force Base, August, 1952.

Manuscript received 5/2/57
Revised manuscript received 7/29/57

PSYCHOMETRIKA—VOL. 23, NO. 1
March, 1958

A RETEST METHOD OF STUDYING PARTIAL KNOWLEDGE AND OTHER FACTORS INFLUENCING ITEM RESPONSE*

VERA T. BROWNLESS AND JOHN A. KEATS†

AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH

A method of studying the problem of correction for guessing and other problems associated with behavior in the test situation is described and an illustrative example presented. As far as the writers are aware this method of approach is novel but, at the same time, it covers many of the practical and theoretical points raised by other writers as reviewed in the introduction.

Awareness of some of the problems involved in tests which are presented in multiple choice form has existed since the early days of testing. One of these problems is based on the fact that the test items can be answered correctly by a person with no knowledge in the field being tested. By purely random selection from the alternatives presented in each question, such a person may obtain a nonzero score on the test. An individual may obtain any score from all correct to none correct, although results for a large group of such persons are expected to yield a group mean which is equal to (total number of questions)/n, where n is the number of choices in each question.

Previous workers attacked this problem in various ways. Many, recognizing that guessing goes on to a greater or lesser degree whatever the instructions, have recommended some form of correction for guessing. In opposition to the idea of making some form of correction, a number of people, in particular Holzinger [4] and Gulliksen [3], have noted that, provided all students answer all questions, the correction factor makes no difference in the rank order of the students. Stanley [7] suggested that although no benefit is derived from the correction when the number of omits varies little from one student to another, the students' attitudes to the testing situation may be improved.

It is doubtful that any over-all guessing correction factor improves the reliability of the test. It is doubtful that all students are guessing from the same number of alternatives; in fact it is quite possible that the more able students can eliminate some of the choices and are therefore guessing among fewer choices. This problem was considered by Horst [5, 6]; he produced a formula that allows for elimination of some choices by some of the students. However, as Davis [1] points out, although this formula allows for partial

*The authors wish to acknowledge help received during discussions with Dr. Frederic M. Lord and Professor S. S. Wilks.

†Present address, Psychology Department, University of Queensland, Australia. The death of Mrs. Vera T. Brownless on May 16, 1957, is regretfully announced.

knowledge it does not make allowance for wrong answers which are based on misinformation. Davis [1] suggests that when a correction formula is used it leads to overcorrection if an examinee has misconceptions, undercorrection if he has partial information, and that these two influences tend to cancel out.

One difficulty in discussing guessing is to find a suitable definition of guessing. In this work the authors are using the one given by Granich "The tendency to answer questions which are unrecognized either wholly or in part, when an answer can not be deduced with certainty from such information as the student possesses" ([2], p. 155). Here no assumption has been made that an n-choice question actually presents n choices to the student. A student with some knowledge may be able to eliminate some choices and thus narrow the field to $n - 1, n - 2, \cdots$, or even 2 choices.

Method of Investigation

To obtain the empirical data for this method it is necessary to administer the *same* test to a group of subjects on two occasions. The time between administrations should preferably be short, and no warning should be given to the subjects that they are going to be retested. If the responses of the subjects to a particular item are examined on the two occasions they will be found to fall into one and only one of the ten categories listed in Table 1. The number of subjects in each category can readily be obtained and these numbers pooled for all items. For example, T_{rr} denotes the number of times any item was marked correctly at both administrations by any subject.

Analysis of the Data

Detailed observation and questioning of subjects while they are taking the tests would probably suggest a large number of factors operating to produce a given response category. For the present, rather simple assumptions will be made, not because they are thought to cover all or even the majority of cases, but to facilitate the description of this method of approach. The possibility of testing these assumptions on the same data should not be overlooked and will be referred to again. It should be noted that the general method of analysis suggested here will not only be useful in investigating the problem of correction for guessing but might well provide an objective method for examining certain factors thought to influence test performance. A simple set of assumptions is given below.

1. At the first administration, all responses are either known correctly, guessed, or "known" incorrectly.
2. At the second administration, all responses are either known correctly, guessed, "known" incorrectly, or repeated from memory.
3. No person who knew the correct answer at the first administration will guess at the second.

TABLE I

Possible Response Categories

Category	Type of Response to an item on two occasions	Number of cases in the category
1	right x right	T_{rr}
2	right x wrong	T_{rw}
3	wrong x right	T_{wr}
4	wrong x same wrong	T_{ww}
5	wrong x different wrong	$T_{w_1w_2}$
6	omit x right	T_{or}
7	omit x wrong	T_{ow}
8	omit x omit	T_{oo}
9	right x omit	T_{ro}
10	wrong x omit	T_{wo}

4. No person will learn an incorrect response between administrations.

The probability that a person who guesses will guess the right answer is regarded as unknown but constant for the persons and items under consideration in the sense that an average figure is required. Obviously subdivisions of items or people or both can be examined separately if sufficient data are available and the corresponding average probabilities for subgroups compared. The problem is to estimate this average probability.

Notation

$1/k$ = the probability of success by guessing.

s = the number of occasions subjects know the correct answer at both administrations.

t = the number of occasions subjects guess at the first administration and know the answer at the second.

u = the number of occasions subjects guess at the same item at both administrations.

m = the number of occasions subjects guess at the first administration and repeat the same response from memory at the second.

x = the number of occasions subjects "know" the same incorrect answer at both administrations.

y = the number of occasions subjects "know" an incorrect answer at the first administration and know the correct answer at the second.

Using this notation as well as that of Table 1 with the assumptions made, it follows that T_{rw}, the number of occasions subjects gave the correct answer on the first occasion and an incorrect answer on the second occasion will equal the product of u, $1/k$, and $(k-1)/k$, when the last term is the probability of guessing a wrong answer the second time. Thus,

$$T_{rw} = \frac{(k-1)u}{k^2}. \tag{1}$$

In a similar way the following four equations can be derived.

$$T_{rr} = s + \frac{t}{k} + \frac{u}{k^2} + \frac{m}{k}. \tag{2}$$

$$T_{wr} = \frac{(k-1)t}{k} + \frac{(k-1)u}{k^2} + y. \tag{3}$$

$$T_{ww} = \frac{(k-1)u}{k^2} + \frac{(k-1)m}{k} + x. \tag{4}$$

$$T_{w_1w_2} = \frac{(k-1)(k-2)u}{k^2}. \tag{5}$$

From (2) and (5), k and u can be estimated.

$$k = \frac{T_{w_1w_2}}{T_{rw}} + 2. \tag{6}$$

$$u = \frac{k^2T_{rw}}{k-1} = \frac{(T_{w_1w_2} + 2T_{rw})^2}{T_{w_1w_2} + T_{rw}}. \tag{7}$$

Although the remaining constants cannot be estimated from the data, it is clear that the difference $T_{wr} - T_{rw}$ is related to the amount of learning during and between the testings, and the difference $T_{ww} - T_{rw}$ is related to the extent of fixation on a particular wrong response. If the material in the test is of an unfamiliar nature it might be safe to assume that there is no prior knowledge and thus that x and y are both zero. In this case s, t and m can be obtained explicitly with the following result:

$$s = T_{rr} - (T_{wr} + T_{ww} - T_{rw})/(k - 1), \tag{8}$$

$$t = (T_{wr} - T_{rw})k/(k - 1), \tag{9}$$

and

$$m = (T_{ww} - T_{rw})k/(k - 1). \tag{10}$$

A second estimate of k can be obtained by considering patterns of responses involving the omission of a response to an item at either or both testings.

Notation

z = the number of occasions a person omits at the first administration and knows the answer at the second.
a = the number of occasions a person omits at the first administration and guesses at the second.
b = the number of occasions a person omits at both administrations.
c = the number of occasions a person guesses at the first administration and omits at the second administration.

A further assumption is required which is in line with the assumptions already listed. It is assumed that persons who know the answer at the first administration will not omit a response at the second administration.

With this assumption and the notation already given, it is possible to derive five more equations in the way illustrated above.

$$T_{0r} = z + \frac{a}{k}. \tag{11}$$

$$T_{0w} = (k - 1)a/k. \tag{12}$$

$$T_{00} = b. \tag{13}$$

$$T_{r0} = c/k. \tag{14}$$

$$T_{w0} = (k - 1)c/k. \tag{15}$$

The solutions for the unknown quantities are as follows:

$$k = \frac{T_{w0}}{T_{r0}} + 1. \tag{16}$$

$$c = T_{r0} + T_{w0} . \tag{17}$$

$$b = T_{00} . \tag{18}$$

$$a = \frac{T_{0w}(T_{w0} + T_{r0})}{T_{w0}}. \tag{19}$$

$$z = T_{0r} - \frac{T_{0w}T_{r0}}{T_{w0}}. \tag{20}$$

With some tests and under certain conditions of administration, the total number of times a person omits an item may be insufficient to give reliable estimates of the constants. In particular, the estimate of k might be based on a relatively small number of cases. This may not be unsatisfactory if this estimate is being calculated only as a check on the value obtained by the method which does not consider omitted items, but it must be noted that in the case of two-choice items this is the only method of estimating k.

Since the primary interest of this type of investigation is the estimation of k, it is important to examine the nature of this estimate. For this purpose consider a person who is guessing between n alternatives for a number of items. Let $k = k_n$, where k_n is the estimate of k obtained from (6).

$$k_n - 2 = T_{w_1w_2}/T_{rw}. \tag{21}$$

This procedure can be repeated for further groups of items provided that within each group the subject is guessing from the same number of alternatives. In practice it is not possible to isolate these groups. The method outlined above yields an average of the following kind:

$$\bar{k} - 2 = \frac{\sum T_{w_1w_2}}{\sum T_{rw}}. \tag{22}$$

It may be difficult to justify this method of averaging over others that suggest themselves in theory. In practice this is the type of average that is given by the present method, and no more satisfactory method has so far been devised for estimating k.

An Illustrative Example

To illustrate the type of results obtained by this approach, data were analyzed for 78 cases from two schools. Each subject had been given two administrations of each of two tests with a period of one week between administrations. The tests used were a mixed verbal and number general ability test (A.C.E.R. Intermediate D) and a nonverbal test involving problems with line figures (Jenkins Test). The frequency of all possible pairs of responses to a given item was tallied, but as there were very few occasions on which an item was omitted, response categories involving an omission are not presented. In Table 2 appear the frequencies for the two tests.

The value of k obtained by applying (22) to these data is 3.6 for Intermediate D and 3.5 for Jenkins Nonverbal. Thus, although these tests both involved five-choice items, the effective number of choices appears to be

TABLE 2

Summed Frequencies in Response Categories for Illustrative Example

	ΣT_{rr}	ΣT_{rw}	ΣT_{wr}	ΣT_{ww}	$\Sigma T_{w_1w_2}$	Total
Intermediate D.	1330	183	285	472	293	2563
Nonverbal	3405	407	939	565	598	5914

about three and one-half as an average over persons and items. A point of contrast between the two tests is suggested by the relatively high value of T_{ww} and low value of T_{wr} for Intermediate D as contrasted with Jenkins Nonverbal. This result suggests that the familiar verbal and number items involved more misconceptions and recall of wrong responses than the unfamiliar items involving classification of line drawings. The latter items, however, showed a greater amount of learning between trials.

It is emphasized that these results are presented to illustrate the method and not to prove anything about the tests. The number of cases is not large and the time between administrations is longer than would ideally be used. However, the results obtained do not appear unreasonable and indicate that further studies of this kind would be worthwhile.

REFERENCES

[1] Davis, F. B. Item analysis in relation to educational and psychological testing. *Psychol. Bull.*, 1952, **49**, 97-121.

[2] Granich, L. A technique for experimentation on guessing in objective tests. *J. educ. Psychol.*, 1931, **22**, 145-156.

[3] Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

[4] Holzinger, K. J. On scoring multiple response tests. *J. educ. Psychol.*, 1924, **15**, 445-447.

[5] Horst, A. P. The chance element in the multiple choice item. *J. gen. Psychol.*, 1932, **6**, 209-211.

[6] Horst, A. P. The difficulty of a multiple choice test item. *J. educ. Psychol.*, 1933, **24**, 229-232.

[7] Stanley, J. C. "Psychological" correction for chance. *J. exp. Educ.*, 1954, **22**, 297-298.

Manuscript received 8/17/56

Revised manuscript received 4/15/57

PSYCHOMETRIKA—VOL. 23, NO. 1
March, 1958

THE MEASUREMENT OF FUNCTION FLUCTUATION

R. F. Garside

UNIVERSITY OF DURHAM, ENGLAND.

A method of measuring function fluctuation is suggested and an appropriate test of significance is indicated. The proposed method is compared with bi-factor analysis and with some other suggested methods of measuring function fluctuation.

The literature on function fluctuations has recently been summarized by Anderson [1]. He considers the various methods which have been proposed and concludes that those suggested by Thouless [12] and Finney [6] not only give similar results but are the best simple methods. Mahmoud ([9], p. 131), however, has stated that Thouless's index of function fluctuation gives results which "seem far too high." Moreover, Finney has intimated [4] that his paper, which Anderson [1] refers to, was a "hurriedly prepared private document" not intended for published discussion.

The accuracy of psychological prediction is limited by the amount of fluctuation in the mental function under investigation. The measurement of such fluctuation is therefore important. Yet it appears that there is no general agreement as to how function fluctuation is best measured—this is the purpose of the present paper.

Definition of Function Fluctuation

Suppose that a group of people are tested on two occasions, that the tests measure a common factor, g, and that the true g scores obtained on each occasion, g_1 and g_2 , are standardized so that the variance of g_1 equals that of g_2 . By fluctuation in function, we mean that the changes in true g scores between occasions $(g_2 - g_1)$ are not constant for all testees. If $(g_2 - g_1)$ is constant, then the function is stable.

Mahmoud ([9], p. 130) refers to such function stability as person stability. It is admitted that person instability is probably a better phrase than function fluctuation, because unequal fluctuations in function is implied rather than fluctuations as such. Nevertheless, the term function fluctuation will be used since it has usually been used in the past to indicate this concept.

Coefficient of Function Stability and of Function Fluctuation

Define the coefficient of function stability, R_{FS} , as the ratio of stable variance in the general factor to variance in a factor general to the same tests given on a single occasion. The coefficient of function fluctuation, R_{FF} , may be defined as $1 - R_{FS}$. Thus,

$$R_{FS} = 1 - R_{FF} = \frac{V_s}{V_{g_1}} = \frac{V_s}{V_{g_2}} , \tag{1}$$

where V_s = the variance of s, the stable part of g_1 and g_2 .

Now suppose that, for each person tested, there is a series of true g scores, each g score being obtained on a different occasion. Then, for a person i

$$g_{ip} = s_i + d_{ip} , \tag{2)}$$

where g_{ip} = g score of person i on occasion p,
s_i = stable score of person i,
d_{ip} = score of person i associated with occasion p.

On each occasion, a set of g scores will be obtained. We may postulate that these sets of g scores are all parallel to each other. Then, if s_i is defined as

$$s_i = \lim_{k \to \infty} \frac{\sum_{p=1}^{k} g_{ip}}{k} , \tag{3}$$

Gulliksen ([7], pp. 28–31) has shown that

$$V_s = r_{gg'} V_g , \tag{4}$$

where V_s = variance of stable scores,
V_g = variance of the set of g scores obtained on any one occasion,
$r_{gg'}$ = correlation between any two such sets of g scores.

Thus, to consider two such sets (or occasions),

$$r_{g_1 g_2} = \frac{V_s}{V_{g_1}} = \frac{V_s}{V_{g_2}} = R_{FS} = 1 - R_{FF} . \tag{5}$$

Neither R_{FS} nor $r_{g_1 g_2}$ can be negative. If $R_{FS} = 0$, $R_{FF} = 1$, and function fluctuation is at a maximum. It should be noted that g_1 and g_2 refer to true g scores. Thus, R_{FF} and R_{FS} are independent of errors of measurement and, therefore, they indicate the extent to which function fluctuation, as such, limits the accuracy of psychological prediction.

In order to measure $r_{g_1 g_2}$, and accordingly R_{FS} and R_{FF} , the plan of using a number of different, not parallel tests, will be adopted. At least two tests must be given on one occasion and at least two other tests on a subsequent occasion. Hence the number of tests must be four or more. No test is given twice, but the same testees take all the tests. This plan differs from that of Thouless [12], who suggests giving two tests twice. It also differs from Dunlap's [3] plan of using four parallel tests given on two occasions.

An essential part of the proposed plan is that the tests must be chosen so that, when *all* the tests are given to a separate group of testees on one occasion, they measure one general factor and no group factors. Whether the intercorrelations so obtained are consistent with this requirement may be ascertained by carrying out a factor analysis or calculating tetrad differences and applying the appropriate tests of significance. An exact test of the significance of tetrad differences has been given by Wishart [14]. In our design, the tests given on the first occasion must not be parallel to those subsequently given, unless all the tests are parallel to each other. Their means, standard deviations, reliability coefficients and specific factor loadings may all differ from test to test.

Strictly speaking, the tests given at the same occasion should be administered simultaneously. This may be achieved by combining the tests into a composite test, each subtest providing items in rotation. It should be remembered, however, that such an arrangement is sound only if the tests are power rather than speed tests. If speed is an important factor, the tests must be given separately.

To simplify the derivation, consider the case when only four tests are used. The derivation may easily be extended to cover five or more tests. If A, B, C, and D represent true scores of the tests and if A and B are obtained at the first occasion and C and D at the second testing then, since the general factor, g, is the sole source of correlation between the tests,

(6) $$r_{AB} = r_{Ag_1} r_{Bg_1}$$

and

(7) $$r_{CD} = r_{Cg_2} r_{Dg_2} \, .$$

But g_i is the sole source of correlation between g_2 and A or B. Therefore

(8) $$r_{AC} = r_{Ag_1} r_{g_1 g_2} r_{Cg_2} \, ,$$

(9) $$r_{AD} = r_{Ag_1} r_{g_1 g_2} r_{Dg_2} \, ,$$

(10) $$r_{BC} = r_{Bg_1} r_{g_1 g_2} r_{Cg_2} \, ,$$

and

(11) $$r_{BD} = r_{Bg_1} r_{g_1 g_2} r_{Dg_2} \, .$$

Substituting (5) in (8), (9), (10), and (11) and multiplying,

(12) $$R_{FS}^4 = \frac{r_{AC} r_{AD} r_{BC} r_{BD}}{r_{Ag_1}^2 r_{Bg_1}^2 r_{Cg_2}^2 r_{Dg_2}^2} .$$

Substituting (6) and (7) in (12),

(13) $$R_{FS}^4 = \frac{r_{AC} r_{AD} r_{BC} r_{BD}}{r_{AB}^2 r_{CD}^2} .$$

Multiplying numerator and denominator of (13) by the variances of A, B, C, and D,

$$R^4_{FS} = \frac{C_{AC}C_{AD}C_{BC}C_{BD}}{C^2_{AB}C^2_{CD}}, \tag{14}$$

where C indicates covariance.

If it is assumed that errors of measurement are uncorrelated with one another or with true scores, then the covariance between the true scores of any two tests equals the covariance between the obtained scores. Thus (14) becomes

$$R_{FS} = \frac{(C_{ac}C_{ad}C_{bc}C_{bd})^{\frac{1}{4}}}{(C_{ab}C_{cd})^{\frac{1}{2}}} = \frac{(r_{ac}r_{ad}r_{bc}r_{bd})^{\frac{1}{4}}}{(r_{ab}r_{cd})^{\frac{1}{2}}}, \tag{15}$$

where a, b, c, and d refer to obtained scores.

Should $r_{ac}r_{ad}r_{bc}r_{bd}$ or $r_{ab}r_{cd}$ be negative, it merely means that the test scores of one or more tests have been inverted. Equation (15) is similar in form to Yule's attenuation formula (Spearman [11], p. 294). The coefficient of function fluctuation, R_{FF}, is given by

$$R_{FF} = \frac{(C_{ab}C_{cd})^{\frac{1}{2}} - (C_{ac}C_{ad}C_{bc}C_{bd})^{\frac{1}{4}}}{(C_{ab}C_{cd})^{\frac{1}{2}}} = \frac{(r_{ab}r_{cd})^{\frac{1}{2}} - (r_{ac}r_{ad}r_{bc}r_{bd})^{\frac{1}{4}}}{(r_{ab}r_{cd})^{\frac{1}{2}}}. \tag{16}$$

If five tests are used a similar derivation gives

$$R^6_{FS} = \frac{r_{13}r_{14}r_{15}r_{23}r_{24}r_{25}}{r^3_{12}r_{34}r_{35}r_{45}}, \tag{17}$$

where tests 1 and 2 are given on the first occasion and tests 3, 4, and 5 on a subsequent occasion. There is no difficulty in deriving R_{FS} for six or more tests.

Mean of R_{FS} and of R_{FF}

The question now arises as to whether $\bar{R}_{FS}$ and $\bar{R}_{FF}$, the mean values obtained from samples, provide unbiased estimates of $\tilde{R}_{FS}$ and $\tilde{R}_{FF}$, the population parameters. Wishart ([14], pp. 184–185) has shown that, when N is large, both $C_{ac}C_{ad}C_{bc}C_{bd}$ and $C_{ab}C_{cd}$ approach the corresponding population parameters. Thus R_{FS} and R_{FF} provide satisfactory estimates of $\tilde{R}_{FS}$ and $\tilde{R}_{FF}$, respectively, when N is large.

Significance of R_{FF}

If the function tested fluctuates between testings, then the intercorrelations between tests will reflect not only a general factor, but also group factors associated with occasions. This was pointed out by Dunlap ([3], p. 448). Thus the significance of R_{FF} may be tested by simply ascertaining the significance of the appropriate tetrad differences in the usual way (Wishart

[14]). When four tests are given, these differences are $r_{ab}r_{cd} - r_{ac}r_{bd}$ and $r_{ab}r_{cd} - r_{ad}r_{bc}$.

It is therefore unnecessary to derive the standard error of R_{FF} or of R_{FS} . If, however, the standard error of R_{FS} is required, it may easily be derived by taking logarithmic differentials (Kelley [8], p. 526) and by using Wishart's [13] moments. These are reported by Kelley ([8], p. 555).

Bi-factor Analysis

It has been suggested that a bi-factor analysis carried out on tests given on different occasions would indicate the extent of function fluctuation. Such an analysis has, in fact, been carried out by Ferguson [5]. He gave three parallel tests to the same group of testees, one test being given on each of three occasions. He then calculated the fifteen correlations between the halves of each test and carried out a bi-factor analysis. He concluded that, "It is not unlikely that both the correlation of errors and functional variability are exerting a positive influence on the size of the group factors, and since no method of determining the relative importance of these two influences is at the moment apparent, it is only possible to describe these factors as factors of temporal contiguity." But when a bi-factor analysis is carried out on correlations among tests designed and administered as described in this paper, then the size of the group factor loadings will be affected only by function fluctuation.

For the sake of simplicity, again consider the case of four tests only, even though this number of tests would be, of course, insufficient to carry out a satisfactory factor analysis. It is assumed that when the four tests are given at the same time, they measure a general factor but no group factors. Thus, when the tests are given in pairs on two different occasions and a bi-factor analysis is carried out, two group factors associated with occasions and a general factor will be obtained.

Note that it is sometimes supposed that it not possible to carry out a bi-factor analysis with two group factors only, unless there is at least one test included which involves neither group factor but the general factor only. But Burt ([2], p. 56) has indicated a method whereby a bi-factor analysis may be carried out when every test has a factor loading on one or the other of the two group factors.

According to our definition of the coefficient of function stability, it equals the ratio of the proportion of test variance attributable to the general factor to the proportion attributable to both general and group factors. If sampling errors are ignored, then this ratio will be constant for all tests, since they measure the same general factor. Thus,

$$R_{FS} = \frac{g_a^2}{g_a^2 + p_a^2} = \frac{g_b^2}{g_b^2 + p_b^2} = \frac{g_c^2}{g_c^2 + q_c^2} = \frac{g_d^2}{g_d^2 + q_d^2}\,, \tag{18}$$

where g_a, g_b, g_c, and g_d are the general factor loadings of the four tests, p_a and p_b are the first group factor loadings, and q_c and q_d the second group factor loadings. Therefore

$$\begin{aligned} R^4_{FS} &= \frac{g_a^2 g_b^2 g_c^2 g_d^2}{(g_a^2 + p_a^2)(g_b^2 + p_b^2)(g_c^2 + q_c^2)(g_d^2 + q_d^2)} \\ &= \frac{g_a^2 g_b^2 g_c^2 g_d^2}{(g_a^2 g_b^2 + g_a^2 p_b^2 + g_b^2 p_a^2 + p_a^2 p_b^2)(g_c^2 g_d^2 + g_c^2 q_d^2 + g_d^2 q_c^2 + q_c^2 q_d^2)} . \end{aligned} \tag{19}$$

But, from (18),

$$g_a p_b = g_b p_a , \tag{20}$$

and therefore,

$$g_a^2 p_b^2 + g_b^2 p_a^2 = 2 g_a g_b p_a p_b . \tag{21}$$

Similarly

$$g_c^2 q_d^2 + g_d^2 q_c^2 = 2 g_c g_d q_c q_d . \tag{22}$$

Substituting (21) and (22) in (19),

$$R^4_{FS} = \frac{g_a^2 g_b^2 g_c^2 g_d^2}{(g_a g_b + p_a p_b)^2 (g_c g_d + q_c q_d)^2} . \tag{23}$$

If scores *a*, *b*, *c*, and *d* are obtained as indicated previously, and if sampling errors are again ignored, then

$$r_{ab} = g_a g_b + p_a p_b , \tag{24}$$

$$r_{cd} = g_c g_d + q_c q_d , \tag{25}$$

$$r_{ac} = g_a g_c , \tag{26}$$

$$r_{ad} = g_a g_d , \tag{27}$$

$$r_{bc} = g_b g_c , \tag{28}$$

and

$$r_{bd} = g_b g_d . \tag{29}$$

Therefore, substituting (24) to (29) in (23),

$$R_{FS} = \frac{(r_{ac} r_{ad} r_{bc} r_{bd})^{\frac{1}{4}}}{(r_{ab} r_{cd})^{\frac{1}{2}}} . \tag{30}$$

Equations (30) and (15) are identical. Therefore, apart from possible differences arising from sampling errors, the method proposed in a preceding section and bi-factor analysis provide equal estimates of the coefficients of function stability and fluctuation. It can be shown, in a similar manner, that this is also true when more than four tests are used. But the proposed method is simpler to carry out.

Comparison with other Coefficients

Paulsen [10] suggested correcting the retest reliability coefficient for attenuation due to test error using the split-half reliability coefficient as the correction factor. The coefficient obtained by this procedure will measure function stability, but Paulsen called it the coefficient of "trait variability." This coefficient is essentially similar to the proposed coefficient, R_{FS} . The proposed coefficient, however, would seem to be superior in that it utilizes more information from the same amount of testing and does not involve the split-half reliability, which does not always provide a satisfactory measure of test error.

Thouless [12] suggests using two tests twice in order to test for and measure function fluctuation. In our notation tests a and c would be the same test administered at different times and so would be tests b and d. Thouless seems to mean the same as we do by function fluctuation and, in fact, points out that if

$$r_{ab}r_{cd} - r_{ad}r_{bc} > 0, \tag{31}$$

then function fluctuation exists. This tetrad difference is the same as one of the pair used in testing for function fluctuation. But Thouless considers that this purpose may be more simply achieved by calculating $r_{(a-c)(b-d)}$. If this correlation is positive, then function fluctuation exists.

To obtain his index of function fluctuation, Thouless divides $r_{(a-c)(b-d)}$ by the mean of r_{ab} and r_{cd} . Accordingly

$$I_{FF} = \frac{2r_{(a-c)(b-d)}}{r_{ab} + r_{cd}}, \tag{32}$$

where I_{FF} is Thouless's index of function fluctuation. Thouless assumes that the standard deviations of a and c and of b and d are equal. He thus obtains

$$I_{FF} = \frac{r_{ab} + r_{cd} - r_{ad} - r_{bc}}{(r_{ab} + r_{cd})\sqrt{(1 - r_{ac})(1 - r_{bd})}}. \tag{33}$$

I_{FF} cannot be directly compared with R_{FF} , since the latter is derived from four separate tests having no group factors. If the same two tests are given twice, (8) and (11) will no longer hold. It is possible, however, to derive a coefficient, R'_{FF} , similar to R_{FF} , using Thouless's experimental design. For (6), (7), (9), and (10) will still apply to the data obtained. Thus R'_{FF} may be derived in a manner similar to that of R_{FF} :

$$R'_{FF} = \frac{(r_{ab}r_{cd})^{\frac{1}{2}} - (r_{ad}r_{bc})^{\frac{1}{2}}}{(r_{ab}r_{cd})^{\frac{1}{2}}}. \tag{34}$$

Apart from the factor $\sqrt{(1 - r_{ac})(1 - r_{bd})}$, I_{FF} only differs from R'_{FF} in that I_{FF} is a function of arithmetic means of pairs of correlation coefficients whereas R'_{FF} is a function of their geometric means. But test *a* is the same as test *c*, and test *b* is the same as test *d*. Therefore, within the limits of sampling error,

(35) $$r_{ab} \simeq r_{cd}$$

and

(36) $$r_{ad} \simeq r_{bc} \; .$$

Thus

(37) $$I_{FF} \sqrt{(1 - r_{ac})(1 - r_{bd})} \simeq R'_{FF} \; .$$

In practice, the factor $\sqrt{(1 - r_{ac})(1 - r_{bd})}$ will be less than unity and will therefore make I_{FF} greater than R'_{FF} ; it appears to be an unnecessary complication. Moreover, as Mahmoud ([9], p. 131) remarks, Thouless's index gives results which seem too high.

Mahmoud [9] considers the case where several tests are given and then repeated in the same or parallel form. He derives a coefficient of person stability, which may be calculated from any number of tests. In the case of two tests only (i.e., four applications) his coefficient reduces to ([9], p. 129, equation xvii)

(38) $$R_{SP} = \frac{r_{ad} + r_{bc}}{r_{ab} + r_{cd}} \, ,$$

where *a* is parallel to *c* and *b* is parallel to *d*. R_{SP} cannot be compared directly with R_{FS} because Mahmoud uses parallel tests. But a coefficient R'_{SF} may be derived, in the same way as R'_{FF} , which will be comparable to R_{SP} ,

(39) $$R'_{FS} = \frac{(r_{ad} r_{bc})^{\frac{1}{2}}}{(r_{ab} r_{cd})^{\frac{1}{2}}} \, .$$

The correlations r_{ad} and r_{bc} , and also r_{ab} and r_{cd} , will again be approximately equal. Therefore R'_{FS} will give similar results to those of R_{SP} , within the limits of sampling errors. The proposed coefficient R_{FS} , however, seems to provide a more direct indication of the extent to which prediction is limited by function fluctuation. Moreover, by avoiding the use of parallel tests, R_{FS} utilizes more information from the same amount of testing than does R_{SP} . It is interesting that giving the same tests twice, or using parallel tests, seems to be a disadvantage in measuring function fluctuation.

Mahmoud ([9], p. 129) states that R_{SP} "measures the extent to which the relative abilities of a given set of persons, assessed on two or more separate days, have remained the same, in spite of the interval between the two applications or (particularly if the interval is short) in spite of the variations in the conditions that obtained." In order, therefore, to obtain a coefficient of trait variability, R_{TV} , Mahmoud subtracts R_{SP} , not from unity, but from his coefficient of internal consistency. This coefficient depends upon errors of measurement, and therefore so does R_{TV} . The proposed coefficient, R_{FS} , is independent of such errors and for our purpose, therefore, would seem to be more appropriate than R_{TV} . It is true that variations in conditions may tend to reduce R_{FS} , but this effect may be minimized by careful test administration.

Example

For an example, some of Mahmoud's data ([9], p. 121, Table II) will be used: $r_{ab} = .713$, $r_{ac} = .881$, $r_{ad} = .637$, $r_{bc} = .559$, $r_{bd} = .670$, and $r_{cd} = .735$ ($N = 87$). From these data, Thouless's coefficient $I_{FF} = .878$. Without the factor $\sqrt{(1 - r_{ac})(1 - r_{bd})}$, I_{FF} would equal .174. It is evident that this factor has a considerable effect, making I_{FF} much greater than R'_{FF} , which equals .176.

Mahmoud's $R_{SP} = .826$, and the coefficient $R'_{FS} = .824$. The results obtained from R_{SP} and R'_{FS} are very similar. The proposed coefficients R_{FS} and R_{FF} , however, include more information from a given amount of testing than does R_{SP} , and their derivation is more direct than that of R_{SP} . Moreover, the proposed coefficients do not entail giving the same tests twice or the use of parallel tests.

REFERENCES

[1] Anderson, C. C. Some simple methods of testing for function fluctuation. *Brit. J. Psychol.*, 1955, **46**, 1-12.

[2] Burt, C. Group factor analysis. *Brit. J. statist. Psychol.*, 1950, **3**, 40-75.

[3] Dunlap, J. W. Comparable tests and reliability. *J. educ. Psychol.*, 1933, **24**, 442-453.

[4] Editorial note. *Brit. J. Psychol.*, 1955, **46**, 230.

[5] Ferguson, G. A. A bi-factor analysis of reliability coefficients. *Brit. J. Psychol.*, 1940, **31**, 172-182.

[6] Finney, D. J. A note on the measurement of performance fluctuation. Memorandum to the National Foundation for Educational Research in England and Wales, 1953.

[7] Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

[8] Kelley, T. L. *Fundamentals of statistics.* Cambridge: Harvard Univ. Press, 1947.

[9] Mahmoud, A. F. Test reliability in terms of factor theory. *Brit. J. statist. Psychol.*, 1955, **8**, 119-135.

[10] Paulsen, G. B. A coefficient of trait variability. *Psychol. Bull.*, 1931, **28**, 218-219.

[11] Spearman, C. Correlation calculated from faulty data. *Brit. J. Psychol.*, 1910, **3**, 271-295.

[12] Thouless, R. H. Test unreliability and function fluctuation. *Brit. J. Psychol.*, 1936, **26**, 325-343.

[13] Wishart, J. The generalized product moment distribution in samples for a normal multivariate population. *Biometrika*, 1928, **20A**, 32-52.

[14] Wishart, J. Sampling errors in the theory of two factors. *Brit. J. Psychol.*, 1928, **19**, 180-187.

Manuscript received 11/19/56

Revised manuscript received 4/1/57

PSYCHOMETRIKA—VOL. 23, NO. 1
MARCH, 1958

PREDETERMINATION OF TEST WEIGHTS

PAUL J. HOFFMAN

THE STATE COLLEGE OF WASHINGTON*

Derivations are presented relating the length of a test to its weight in a composite. Tests of varying length are constructed so that their weights will be of predetermined magnitudes, and the results compared with expectations. Weighting schemes involving standard deviations of raw scores and of true scores are compared. An important secondary derivation is presented from which it is possible to estimate test reliability knowing only the relative length of a test, its shortened form, and the standard deviation of each.

Given test A with known variance and reliability, one frequently wishes to construct a second test, B, such that the relative weights of the two tests for determining a composite score will be of some predetermined magnitude. Where test B can be experimentally pretested, item analysis procedures designed to control the standard deviation and reliability of the test can be applied ([1], pp. 375–380). If item parameters cannot be obtained in advance, the usual practice is to construct test B without regard to the problem of weighting and to apply some transformation to the scores after the test is administered and the test parameters determined.

In many applications, and particularly in the classroom, the person responsible for evaluation is not prepared to engage in what seems to him to be high-powered statistical manipulations. What is wanted is a way of arriving at a composite for each individual member of his class by simply totaling the various part scores. For this reason, an attempt is often made to predetermine weights by controlling the number of items in each test. It has been shown ([1], pp. 336–341) that the number of items in a test is not a necessary determinant of test weight, a fact which might appear to rule out this possibility as a solution. It is not known, however, precisely how the number of items is likely to affect test weights. Since practical people may well continue to justify its use in the lack of strong evidence to the contrary, it becomes important to determine the conditions under which weighting by controlling the number of items in a test may be successfully employed, and the conditions under which it may not.

The matter is somewhat complicated since the concept of test weight is itself not clearly defined. There are a variety of suggestions for *equalizing* the contributions of two or more tests in the absence of a criterion ([3], pp. 211–213; [4], pp. 88–90) and some suggestions for determining whether a given test contributes more than or less than another [5]. Each method implies

*Now at University of Oregon.

a somewhat unique definition of *weight*. It is not our purpose to re-examine the problem of the meaning of test weights. Instead, we consider two definitions of test weight and develop the methods for their predetermination on the basis of length of test.

Weighting by Standard Deviations

It is often assumed that the effective weight of a test in relation to another is determined by the ratio of the standard deviations of the two tests. Thus, if test X has a standard deviation σ_x and test Y a standard deviation σ_y , the weight of Y in relation to X is given by $W_y = \sigma_y/\sigma_x$.

Now let us assume X is a test of unit length, and that Y is a test of increased length, such that, in deviation scores, $y = x_1 + x_2 + \cdots + x_k$. Then

$$\begin{aligned} \sigma_y^2 &= \frac{\sum (x_1 + x_2 + \cdots + x_k)^2}{N} \\ &= \sum_{i=1}^{k} \sigma_{x_i}^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} r_{x_{ij}} \sigma_{x_i} \sigma_{x_j} , \qquad (i \neq j). \end{aligned} \tag{1}$$

If it is assumed that the components of Y are parallel forms, one may substitute as follows:

$$\sigma_{x_i}^2 = \sigma_x^2 ; \qquad r_{x_i x_j} = r_{xx} ,$$

so that from (1),

$$\sigma_y^2 = k\sigma_x^2 + k(k-1)r_{xx}\sigma_x^2 .$$

But

$$W_y^2 = \frac{\sigma_y^2}{\sigma_x^2} = \frac{k\sigma_x^2 + k(k-1)r_{xx}\sigma_x^2}{\sigma_x^2} = k + k(k-1)r_{xx} .$$

Therefore,

$$W_y = \sqrt{k + k(k-1)r_{xx}} . \tag{2}$$

From (2) it is seen that the effective weight of a test varies directly as a function of test length and reliability. If the reliability of the unit test is 1.00,

$$\begin{aligned} W_y &= \sqrt{k + k(k-1)} \\ &= \sqrt{k + k^2 - k} ; \\ W_y &= k. \end{aligned} \tag{3}$$

If the reliability of the unit test is zero,

$$W_y = \sqrt{k}. \tag{4}$$

Considering (3) and (4), the inequality

$$\sqrt{k} \leq W_y \leq k$$

makes the dependence of test weight upon length obvious.

Our main concern, however, is that of finding a value for k that will result in a predetermined weight W_y . To solve (2) for k, first square both sides:

$$\begin{aligned} W_y^2 &= k + k(k - 1)r_{xx} \\ &= k + k^2 r_{xx} - k r_{xx} \, . \end{aligned}$$

Arranging terms in quadratic form,

$$r_{xx}k^2 + (1 - r_{xx})k - W_y^2 = 0;$$

(6) $$k = \frac{-(1 - r_{xx}) \pm \sqrt{(1 - r_{xx})^2 + 4r_{xx}W_y^2}}{2r_{xx}}.$$

Since a negative radical leads to $k < 0$, only one root is meaningful:

(7) $$k = \frac{\sqrt{(1 - r_{xx})^2 + 4r_{xx}W_y^2} - (1 - r_{xx})}{2r_{xx}}.$$

From (7), one can estimate the relative length of a test that is required in order to yield a given weight with respect to the unit test.

Example: Assume that the cumulated scores for an individual to the end of the semester comprise a total of 100 test items. The reliability of the cumulation is .70. It is desired to construct a final examination which will equal twice the weight of the other tests. In this example,

$$r_{xx} = .70, \qquad W_y^2 = 4.00,$$

$$k = \frac{\sqrt{(.30)^2 + (4)(.70)(4.00)} - .30}{(2)(.70)} = \frac{\sqrt{11.29} - .30}{1.40};$$

$$k = 2.19.$$

Then the number of items necessary on the final examination is given by $100k = (100)(2.19) = 219$ items.

(It should be noted that the terms of (6) can be rearranged to yield an expression for r_{xx} in terms of W_y and k. Thus,

$$r_{xx}(k^2 - k) = W_y^2 - k;$$

$$r_{xx} = \frac{W_y^2 - k}{k^2 - k}.$$

This formula for reliability of a shortened form of a test requires only the standard deviation of the initial test, the standard deviation of the shortened form, and their relative length.)

Figure 1 is a nomograph from which k can be quickly determined for any given r_{xx} and any desired W_y .

It should be emphasized that the derivation of W_y depends upon one important assumption: that the components of Y are parallel forms of test X. For the development of aptitude tests this may impose no significant practical

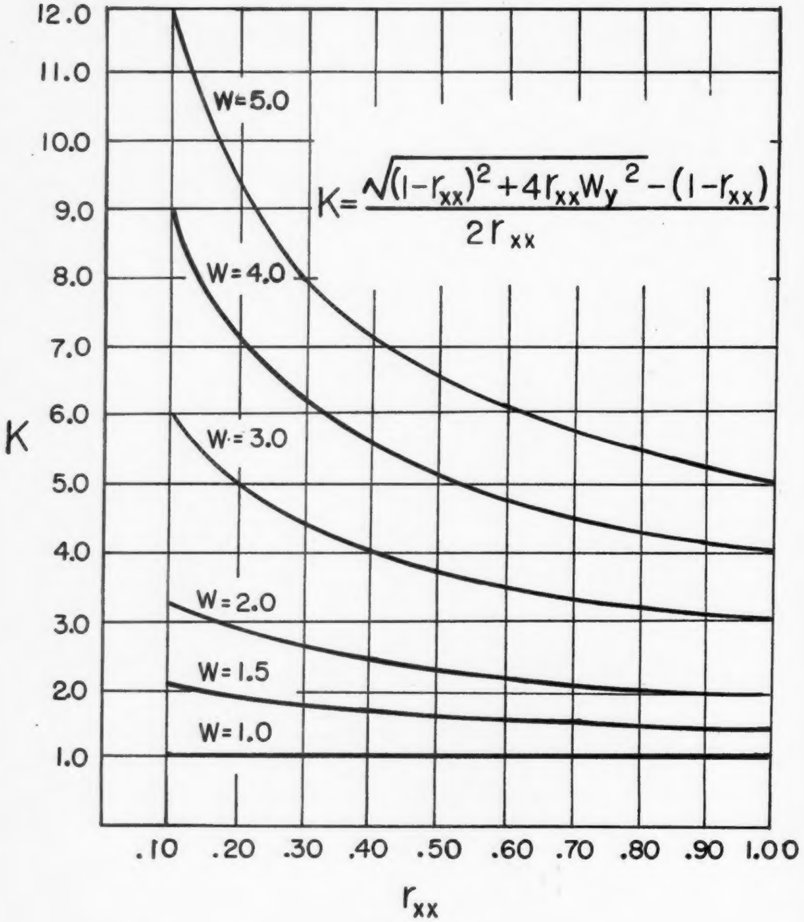


FIGURE 1

Computing diagram for estimating the length of a test, Y, such that $W_y = \sigma_y/\sigma_x$, where:

W_y = the desired weight of the test Y,
r_{xx} = reliability of test X,
K = ratio of estimated length of Y to length of X.

limitation, but for achievement testing the situation is different. Achievement testing at different stages of learning yields scores on individuals who may differ in their rate of learning. In addition, course content is not necessarily highly interdependent among its various stages. For these reasons it seems reasonable to doubt the comparability of two achievement tests separated by a period of learning, unless some empirical evidence can be offered to show that such a procedure makes little practical difference. We shall return to the empirical question in a later portion of the paper.

Weighting by True Scores

One major difficulty in assuming that the weight of a test is a function of its standard deviation is that tests of low reliability will necessarily have small standard deviations. Thus, scores of an unreliable test may be multiplied by a constant so as to increase the test's standard deviation in relation to a second more reliable test. The composite score thus becomes contingent upon the more unreliable test. This difficulty has been acknowledged, ([2], pp. 385–396) but the proposals for overcoming it have been varied. A solution that meets this objection, and one which seems to make rational sense is to define test weight in terms of the ratio of the standard deviations of true scores. Thus,

$$W_y = \sigma_{t_y}/\sigma_{t_x} . \tag{8}$$

In what manner does test length affect test weight defined in this way? Let us regard the true score on test Y as composed of tests of unit length, X. In deviation scores,

$$t_y = \sum_{i=1}^{k} t_{x_i} . \tag{9}$$

Again assuming comparable forms among the components of Y, it follows that the t_{x_i} will be equal. Then (9) becomes

$$t_y = kt_x ,$$

and

$$\sigma^2_{t_y} = k^2 \sum t_x^2/N = k^2\sigma^2_{t_x} .$$

Solving for k,

$$k^2 = \sigma^2_{t_y}/\sigma^2_{t_x} ; \qquad k = \sigma_{t_y}/\sigma_{t_x} .$$

Substituting from (8),

$$k = W_y . \tag{10}$$

Equation (10) states that if test weight is defined as the ratio of the sigmas of true scores, increasing the length of the test by the proportion k

increases its weight by k also. Thus, if one wishes to write a test that will count twice as much as a given test, he simply writes twice the number of items. This coincides with the intuitively justified practices of many teachers who have no knowledge of test theory. The practice can now be seen to be statistically justified, when the assumption of parallel forms is met.

To obtain evidence concerning the accuracy with which estimates of W_y can be made, scores were obtained from midsemester and final examinations in Introductory Psychology for a group of 54 college freshmen. Both examinations were multiple choice, the final consisting of 105 items. Only the first 30 items of the midsemester examination were used. Successive portions of the final examination were scored, yielding totals for each individual for the first 30, 60, 75, 90, and 105 items. The successive scores are thus not independent, a fact which detracts from the meaningfulness of the comparisons but which does not invalidate them. These results are plotted in Figures 2, 3, and 4. Figure 2 compares obtained values, $W_y = \sigma_y/\sigma_x$, with values of W_y estimated from (2). In this case, X is the 30-item midsemester examination, and Y is the final exam of varying length. Figure 3 differs from Figure 2 only in that test X now consists of the first 30 items of the final examination,

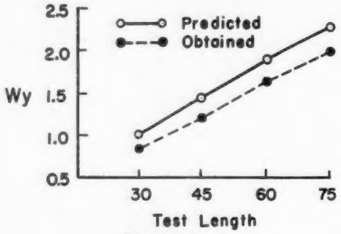


FIGURE 2

Predicted and obtained weights of test Y. Predictions made on the basis of 30-item midterm examination. Test Y consists of accumulations of items of the final examination beyond the first thirty.

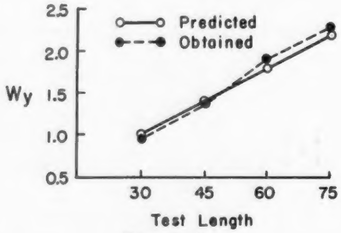


FIGURE 3

Predicted and obtained weights of test Y. Predictions made on the basis of first 30 items of final examination. Test Y consists of accumulations of items of the final examination beyond the first thirty.

and test Y is composed of the successive portions of this examination, not including the first 30 items. It can be assumed that the difference between these two figures is due to the fact that the assumption of comparable forms is more nearly met for the latter situation than for the former.

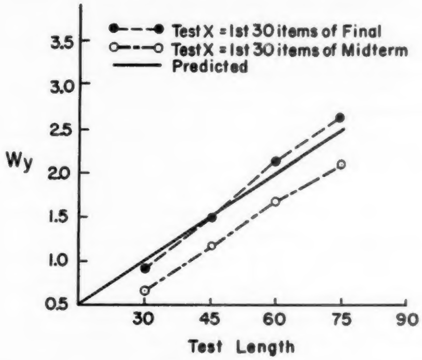


FIGURE 4

Predicted and obtained weights of test Y. Test weight defined as ratio of true scores.

Figure 4 shows the predicted and obtained values of W_y when defined as a ratio of true scores, according to (8). In this case, the predicted values are exactly proportional to test length; hence the solid diagonal line represents these predictions. The actual obtained values for W_y were in this case determined by noting that $\sigma_{t_y}^2 = r_{yy}\sigma_y^2$ and $\sigma_{t_x}^2 = r_{xx}\sigma_x^2$. Therefore,

$$W_y = \frac{\sigma_y \sqrt{r_{yy}}}{\sigma_x \sqrt{r_{xx}}}.$$

The reliabilities were estimated from the item data, using Kuder-Richardson Formula 20. As was apparent in a comparison of Figures 2 and 3, so too in Figure 4, the use of the first 30 items of the final examination as test X results in predictions which appear to be more accurate than those based on the midsemester examination. The necessity for satisfying the assumption of parallel forms seems again to be indicated.

It should be emphasized that the definition of parallel forms, necessary to satisfy the assumptions of the equations derived in this paper, is one which demands only that the variances and intercorrelations be equal. We need not say that the intercorrelations are perfect, or even that they are high. To assume such identity would reduce the entire question of differential weighting to a triviality, except as it may lead to the maximization of the reliability of the composite or to the prediction of an external criterion.

REFERENCES

[1] Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

[2] Horst, P. The prediction of personal adjustment. *SSRC Bulletin No. 48,* 1941.

[3] Kelley, T. L. *Interpretation of educational measurements.* New York: World Book, 1927.

[4] Thurstone, L. L. *The reliability and validity of tests.* Ann Arbor, Michigan: Edwards Bros., 1931.

[5] Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika,* 1938, **3**, 23-40.

Manuscript received 7/23/56

Revised manuscript received 4/18/57

RULES FOR PREPARATION OF MANUSCRIPTS FOR *PSYCHOMETRIKA*

1. Send manuscripts to the Managing Editor:

   LYLE V. JONES
   Psychometric Laboratory
   University of North Carolina
   Chapel Hill, North Carolina

2. Submit three typewritten copies of the manuscript. For original copy use heavy white typewriter paper, size 8½ x 11. Double-space the lines, leave ample space around formulas, and allow wide margins for editorial work.
3. Accompanying the manuscript should be three copies of an Abstract of no more than 100 words, outlining the contents of the paper.
4. Tables should be submitted with the manuscript in four copies. Prepare original copy of tables on electric typewriter, in a form suitable for photographic reproduction. The remaining three copies need not be prepared on an electric typewriter, but should adhere to the prescribed form.

   Tables are to be numbered with Arabic numerals and referred to in the text by number' e.g., Table 2. The heading of the table should be centered. The word "Table," on the first line of the heading, should appear in capital letters, e.g., TABLE 2. The title, double-spaced below the table number, should have initial letters of principal words capitalized. Titles should be short; if two lines are required they should be single-spaced.

   Double horizontal lines should separate the heading from the stubhead, a single line should appear between the stubhead and the body of the table, and a single line should appear at the bottom of the table. Footnotes referring to any part of the table should be single-spaced immediately below the table. Tables appearing in *Psychometrika*, 1956, 21, 362-363 show a variety of examples in good form.

   For the electrically typed copy of tables, heavy white paper should be used, and no erasures should appear. Corrected entries may be pasted over errors using rubber cement. On this copy, closely related tables should be prepared or mounted on the same sheet in such a way that final copy will fit the journal page after reduction. If this results in a sheet size exceeding 8½ x 11 inches, the use of mailing tubes is recommended.
5. Figures should be drawn only by an expert draftsman, about three times the size at which they will appear. They should be on plain white paper or tracing cloth in black India ink. They should be referred to in the text by number, e.g., Fig. 3. Each figure caption, including the figure number and a succinct title, should be typed on a separate sheet of paper. No such identification should appear on the front of the figure. On the margin of the back of the figure the author should write lightly his name and the figure number. In addition to the original copy of figures, three photographic reproductions or rough sketches of figures should be submitted with the manuscript.
6. Formulas should be numbered at the left margin with Arabic numerals in parentheses. Careful attention should be given to the punctuation of formulas, which ordinarily are to be regarded as parts of sentences. Formulas should be legible, and unfamiliar symbols avoided if possible. Where they are used for the first time, they should be defined in the margin, as "upper case Greek letter gamma." For very complicated notations, a list for the use of the printer should be submitted.

7. Footnotes to the text should be reduced to a minimum. Formulas in footnotes should be avoided. Footnotes should be indicated by the following symbols: *(asterisk), †(dagger), ‡(double dagger), §(section mark), ||(parallels), ¶(paragraph mark). Footnotes should be typed at the bottom of the page of text to which they refer.

8. References should be segregated at the end of the article. The heading should be "References" not "Bibliography," and should be capitalized and centered. The references in such a list should be arranged in alphabetical order according to author's name, and numbered with Arabic numerals in brackets. In the text references and pages should be referred to by number: [2], [2, 6, 10], [cf. 3], [e.g., 4, 6], ([2], p. 36), (cf. [2], [5], p. 20, eq. 13).

   With only minor exceptions, the forms of citation adopted by the Board of Editors of The American Psychological Association are used in *Psychometrika*. (See American Psychological Association, Council of Editors. *Publication manual of the American Psychological Association, 1957 revision*. Washington, D. C.: American Psychological Association, 1957.) The form for a journal reference is as follows:

   [1] Gulliksen, H. and Tucker, L. R. A mechanical model illustrating the scatter diagram with oblique test vectors. *Psychometrika*, 1951, **16**, 233-238.

   The form for a book reference is as follows:

   [6] Thurstone, L. L. *Multiple-factor analysis*. Chicago: Univ. Chicago Press, 1947.

9. A separate sheet giving the title of the article and the author's name and professional connection should be included with the manuscript. The author's name or professional connection should not appear on the manuscript. There should be no reference which would identify the author of the manuscript, e.g., "In previous work [14], the present writer has shown that . . ." Since all such statements must be removed before the article is sent to the editors, it will facilitate work in the editorial office if these precautions are observed.

10. The author is urged to give careful attention to grammatical construction, spelling, and punctuation

11. The journal will provide 100 free offprints of each article. Additional offprints will be available in accordance with the following schedule:

	2 pp.	4 pp.	8 pp.	12 pp.	16 pp.	Add. 2 pp.
100 copies	$4.00	$8.00	$12.00	$16.00	$20.00	$2.00
Each additional 100	$2.00	$4.00	$ 6.00	$ 8.00	$10.00	$1.00

   A blank page counts as one page. Covers: $12.00 first hundred, $5.00 each additional hundred.